THE PHYSICAL SOCIETY

REPORTS ON PROGRESS IN PHYSICS

VOLUME XI (1946–47)

Price to non-Fellows £2 2*s. net* (*Postage* 1*s.*)
Price to Fellows 25*s. net* (*Postage* 1*s.*)

Published by
THE PHYSICAL SOCIETY
1 Lowther Gardens, Prince Consort Road
London S.W.7
1948

PRINTED IN GREAT BRITAIN

CONTENTS

PAGE

Preface iv

Electrostatic generators for the acceleration of charged particles. By R. J. Van de Graaff, J. G. Trump and W. W. Buechner . . . 1

Radioactive branching. By N. Feather, F.R.S. 19

The neutrino and the recoil of nuclei in beta disintegrations. By B. Pontecorvo 32

Ferromagnetism. By Edmund C. Stoner, F.R.S. 43

The calculation of atomic structures. By D. R. Hartree, F.R.S. . . 113

Developments in the infra-red region of the spectrum. By G. B. B. M. Sutherland and E. Lee 144

The radio-frequency spectroscopy of gases. By B. Bleaney . . . 178

Ultrasonics research and the properties of matter. By Charles Kittel . 205

Latent image formation in photographic silver halide gelatine emulsions. By W. F. Berg 248

The mechanism of the thermionic emission from oxide coated cathodes. By H. Friedenstein, S. L. Martin and G. L. Munday 298

Physics in glass technology. By H. Moore 342

Evaporation in nature. By H. L. Penman 366

Meteors, comets and meteoric ionization. By A. C. B. Lovell, J. P. M. Prentice, J. G. Porter, R. W. B. Pearse, N. Herlofson:

- Meteors. By J. P. M. Prentice 389
- Comets and meteor-streams. By J. G. Porter 402
- Spectroscopic observations on comets, meteors and meteorites. By R. W. B. Pearse 410
- Meteoric ionization and ionospheric abnormalities. By A. C. B. Lovell 415
- The theory of meteor ionization. By N. Herlofson . . . 444

Author index 455

PREFACE

In previous years it has been the custom for a general editor to take the responsibility for the production of these reports. In view of the great increase in work in universities and other institutions immediately following the end of the war, it was decided to constitute a small editorial board to supervise the present volume. The members of this Board were Professor R. E. Peierls, F.R.S., Mr. C. H. Collie, Dr. W. Jevons, Dr. N. Kemmer and Mr. T. Smith, F.R.S., with the writer as Chairman. The Board are extremely grateful to Dr. W. B. Mann, who edited Vols. VIII, IX and X, for the substantial help he has given in the preparation of Vol. XI, in addition to his invaluable guidance.

The Board regret that the present volume must again be regarded as covering a period of two years, but now see no reason why the pre-war custom of an annual volume cannot be resumed as from 1948.

The Board wish to express their thanks to Dr. F. Lowater for preparing the index and to Dr. A. C. Stickland, on whom has fallen the real burden of most of the detailed work.

G. B. B. M. SUTHERLAND,
Chairman.

AUTHOR INDEX

Abegg, R., 282.
Abney, W. de W., 255, 256, 279.
Adams, N. I., Jr., 20.
Adel, A., 192.
Admiralty Research Laboratory Reports E 300, 149, 152, 154, 159, 160, 169.
Aiken, C. B., 163.
Akhieser, A., 232.
Albert, E., 254.
Albrecht, E., 20.
Albricht, W., 310, 311.
Alexander, E. A., 215.
Allan, G. E., 233.
Allekotte, —, 155, 156, 159.
Alleman, R. S., 213.
Allen, J. S., 35, 37, 38, 39.
Allison, H. W., 341.
Anderson, H., 271.
Andrews, D. H., 164.
Andrianov, A. M., 331.
Andrias, J., 9.
Angot, A., 381.
Ångström, A., 376.
Anselm, A., 221, 226.
Appleton, E. V., 415, 416, 419–424, 426–429, 431, 438, 441, 444.
Aranovitch, R. M., 335.
Arenberg, D. L., 226, 233, 234, 237.
Arens, H., 279.
Arnot, F. L., 125.
Arrol, W. J., 414.
Ars'eneva, A. N., 315.
Ashby, R. M., 16.
Aten, A. H. W., Jr., 306.
Autler, S. H., 183, 187, 191, 201.
von Auwers, O., 95, 97, 104, 109.
Awbery, J. H., 82.

Babcock, H. D., 193.
Babcock, H. W., 411.
Baber, T. D. H., 132, 133.
Baccaredda, M., 225.
Backhouse, T. W., 392.
Bagdassarian, C., 264, 285.
Baker, C. J. W., 99.
Band, W., 74.
Banerji, S. K., 378.
Banwell, C. J., 428–431, 434, 437–440, 452, 453.
Bär, R., 226.
Bardeen, J., 199, 319.
Barker, E. F., 196.
Barnes, R. B., 174, 175.
Barnet, C. E., 147.
Barnett, M. A. F., 415.
Barone, 225.
Barr, E. E., 163, 165.
Barrer, R. M., 320, 321.
Barshchevsky, —, 262.
Bartelt, O., 257.
Bartlett, J. H., 116, 131, 132.
Barton, H. A., 7.
Bates, D. R., 129, 140, 444, 448.
Bates, L. F., 77, 94, 99, 100, 109.
Batschinski, A. J., 226.
Bauer, C. A., 414.
von Baumbach, H. H., 315.
Baur, E., 256.
Baxter, S., 258, 271.
Bazulin, P. A., 225.
Becker, G., 93.
Becker, G. E., 183, 187, 191.
Becker, J. A., 163, 301, 305, 308, 310, 312, 321, 322.
Becker, R., 108, 228, 232.
Beckerath, K., 266.
Beckett, S. H., 383.
Beeman, W. W., 93, 94.
Beers, Y., 200.
Bell, E. E., 164, 174.
de Benedetti, S., 22.
Benedict, D. L., 15.
Benfield, A. E., 237.
Benjamin, M., 302, 304, 318.
Bennet, W. E., 16.
Berdennikowa, T. P., 302, 312, 315.
Berg, O. E., 171.
Berg, W. F., 248, 249, 252, 253, 260, 261, 263, 264, 271, 272, 274–280, 288–292.
Bergmann, L., 237.
Beringer, R., 182, 185, 191, 194.
Berkner, L. V., 419, 421, 423, 427.
Bernert, T., 23, 25.
Bernet, E. J., 15.
Berriman, R. W., 293.
Berthelot, —, 23.
Best, A. C., 370.
Best, J. E., 419, 423.
Bethe, H. A., 42, 128.
Beuttenmüller, E., 226.
Beverage, H. H., 442*.
Beyer, R. T., 225.
Bhagavantam, S., 237.
Bhar, J. N., 419, 421, 423, 427.
Biermann, L., 129.
Bigelow, F. H., 378, 385.
Bilham, E. G., 381.
Billings, B. H., 163, 165.
Biltz, M., 265, 269.
B.I.O.S., 149.
Biot, E. C., 391.
Biquard, P., 227.
Bittencourt, P. T., 22.
Bitter, F., 100, 108.
Black, M. M., 124.
Blackett, P. M. S., 436, 452.
Blackwell, D. E., 159, 160, 172, 174.
Blau, M., 293.
Bleaney, B., 178, 181, 183, 185, 186, 190, 196, 199–202.
Bleuler, E., 30.
Bleuler, K., 233.
Blewett, J. P., 300, 302, 306–308, 310, 311, 315, 317, 320, 321, 323, 326.
Bloch, F., 35, 37.
Bloch, O., 268.
Blout, E. R., 169.
Bobrovnikoff, N. T., 411.
Bodenstein, M., 257.
de Boer, J. H., 257, 300, 306, 307, 312, 314–316, 319–321, 325–327, 335.
Bole, G. A., 238.
Bömmel, H., 237.
Bommer, H., 85.
Bonner, T. W., 16.
Boothroyd, S. L., 397.
Boot, A. H., 331, 335.
Born, H., 237.
Borzyak, P., 329.
Bose, J. C., 158.
Bowden, B. V., 20–22.
Bowden, F. P., 172.
Bowen, E. G., 419, 424.
Bowen, I. S., 377.
Bozorth, R. M., 109.
Bradley, A. J., 106.
Bradt, H., 26, 28.
Brattain, W. H., 163, 315.
Breido, I. I., 252, 265.
Breit, G., 116, 415.
Brentano, J. C. M., 258, 271, 293.
Bretscher, E., 26.
Briggs, L. J., 368.
Brillouin, L., 223, 224.
Broda, E., 23.
Broili, H., 293.
Bronson, H. L., 20.
Bronstein, M., 312.
Brooker, L. G. S., 267.
Brown, F. W., 122.
Brown, H. T., 368.
Bruining, H., 306, 335.
Brunt, D., 373, 375, 376.
Buckingham, E., 383.
Buechner, W. W., 1, 8, 13.
Builder, G., 420.
Bullock, E. R., 254, 256.
Burgers, W., 307, 319.
Burrill, E. A., 1, 8, 13.
Burrows, R. S., 255.
Burton, P. C., 252, 253, 275–277, 290.
Byers, W. H., 215.

Cabannes, J., 276.
Cady, W. G., 237.
Caldwell, S. H., 127.
Cameron, A. E., 253.
Cannon, W. F., 202.
Carey Lea, M., 256.
Carlson, J. F., 21.
Carpenter, L. G., 367, 378.
Carrier, W. H., 378.

* Not referred to in the text.

Carroll, B. H., 254, 257, 259, 261, 262, 267, 268, 273, 274.
Carter, R. L., 202.
Carter, W. H., 163.
Cartwright, C. H., 160.
Case, T. W., 151.
Cashman, R. J., 151, 152, 154, 172.
Cefola, M., 225, 234, 237.
Chadwick, J., 32.
Chalfin, E., 22.
Chalmers, T. A., 22.
Chamanlal, —, 441.
Chandrasekhar, S., 139, 140.
Charlton, E. E., 7.
Chick, A. J., 238.
Chilton, L. V., 293.
Christiansen, B. J., 164.
Claassens, A., 307.
Clark, C. W., 82.
Clark, J. R., 99, 100.
Clark, W., 253, 280, 281.
Clausing, P., 310, 329, 336.
Cleeton, C. E., 180, 195.
Clegg, J. A., 428, 429, 431, 432, 436, 437, 439, 440, 452, 453.
Clerke, A. M., 393.
Clogston, A. M., 130.
Cloud, R. W., 1, 7.
Coeterier, F., 52.
Coles, D. K., 183, 196–198, 200.
Compton, K. T., 1, 308, 446.
Condon, E. U., 210.
Condon, E. V., 113, 115, 120, 128, 129, 143.
Coolidge, A. S., 116, 129, 131, 132.
Coomes, E. A., 331, 332, 334.
Cooper, F. S., 7.
Corey, V. B., 225.
Corry, J. E., 225.
Corson, D. R., 25.
Coulson, C. A., 130.
Cox, J. W., 427.
Crane, H. R., 36, 37.
Crawford, F. W., 175.
Cross, P. C., 193, 199.
Crowther, E. M., 383.
Cüer, P., 293.
Cummings, N. W., 376, 377.
Curtis, R. W., 216.
Czerlinsky, E., 233.
Czerny, M., 166.

Dacus, E. N., 161, 162, 174.
Dahl, O., 1, 3.
Dailey, B. P., 197, 199, 200.
Dakin, T. W., 198.
Dalton, J., 366, 367.
Daly, E. F., 148, 173.
Dannöhl, W., 108.
Darbyshire, J. A., 307, 319, 321, 328.
Darwin, C. G., 19.
Daubree, G. A., 367.
Daunt, J. G., 162, 164.
Davidson, J. M., 367.
Davidson, M., 398.
Davisson, C., 308, 322.
Davydov, B., 317.
Davydov, V. K., 380.
Debot, R., 266, 278–280.
Debye, P., 180, 188, 227.
Dehlinger, U., 77, 109.
Delano, R. B., Jr., 238.
Delbrück, M., 114.
Dellinger, J. H., 443*.
Demers, P., 293.
Denning, W. F., 393, 398.
Dennison, D. M., 191, 195–197.
Derenzini, T., 225.
Dersch, F., 252, 253.
Detels, F., 301, 303.
D'Eustachio, D., 236.
Deutsch, M., 28, 29.
Deutsch, M. L., 200.
Dicke, R. H., 178, 184, 185, 191, 192.
Dieke, G. H., 193.
Dirac, P. A. M., 117.
Dobson, G. M. B., 444.
Döring, W., 108, 228, 232.
Dreibelbis, F. R.,
Droz, M. E., 225, 234.
Dubar, W. S., 225.
Dubiago, A. D., 404.
Dumond, J., 131.
Duncanson, W. E., 130.
Dunn, C. G., 116, 132.
Dünwald, H., 315.
Dunworth, J. V., 26.
Dürr, H., 252, 253.
Dutta, A. K., 225.
Duyckaerts, G., 82.

Eastwood, E., 428.
Ebermayer, E., 368.
Eckersley, T. L., 416, 422–428, 435, 438.
Eckhart, C., 226.
Edlén, B., 139.
Edwards, O. S., 84.
Eggert, J., 259, 265–267, 269, 271, 284, 293.
Ehret, L., 237.
Eisenstein, A., 302, 318, 319, 330–334.
Ellickson, R. T., 171.
Elliot, A., 199.
Elliott, A., 160.
Ellis, C. D., 32, 33.
Ellyett, C. D., 428.
Elvey, C. T., 411.
Elyash, E. S., 200.
Emslie, A. G., 237.
English, —, 23.
Epstein, L. F., 237.
Epstein, P. S., 444, 445.
Eriksson, A., 129, 132.
Ernst, P. J., 237.
Escombe, F., 368.
Eser, C., 368.
Espe, W., 301–305, 312.
Evans, C. H., 253, 268, 276, 277, 279, 289, 291, 292.
Evans, R. M., 256.
Evans, W. E., 441, 442.
Ewald, P. P., 370, 371.
Ewles, J., 164.
Eyring, H., 218, 219, 221.

Fairbank, H. A., 237, 240.
Fairbank, W. M., 237, 240.
Fajans, K., 19, 20, 23, 266.
Falkenhagen, H., 237.
Falla, L., 280.
Fallot, M., 45, 67, 83, 86, 87, 89, 96, 98, 100, 103.
Fan, H. Y., 304, 306, 317, 327.
Farcas, T., 87.
Farineau, J., 93.
Farmer, F. T., 419, 423, 426.
Farrer, W. J. G., 256.
Fay, C. H., 67.
Feather, N., 19, 21, 23, 24, 26.
Federal Communications Commission, 426, 427, 438.
Feldenkrais, M., 7.
Feldman, W., 315.
Fellgett, P. B., 172.
Fermi, E., 33, 37, 41.
Ferrell, O. P., 428.
Ferry, J. D., 223.
Fesefeldt, H., 264.
Feshbach, H., 1.
Fessenkoff, B., 407.
Fineman, A., 318, 330–333.
Firestone, F. A., 160, 238.
Fischer, F., 158.
Fisher, W. J., 398.
Fitzgerald, D., 378.
Fitzgerald, J. W., 236.
Flammersfeld, A., 30.
Flory, L. E., 149.
Flügge, S., 23.
Fock, V., 117, 122, 123, 129, 140.
Follansbee, R., 377, 381.
Folse, J. A., 377.
Fonda, G. R., 147.
Fofrer, R., 45, 92.
Fortier, S., 383.
Foster, J. F., 449.
Fowler, R. H., 109, 312–315, 317, 324, 329.
Fox, F. E., 225, 227, 238.
Franck, J., 269, 326.
Frank, H. S., 218.
Frank, N. H., 128.
Frankel, S., 225, 234.
Frankenburger, W., 266.
Frederick, J. R., 238.
Freeman, J. R., 378.
Frei, H., 225.
Frenkel, J., 228.
Freundlich, H., 253.
Fricke, E. F., 212, 213, 217.
Friedmann, H., 93.
Friedrich, E., 94, 101.
Friedrich, W., 293.
Frisch, D. H., 16.
Fritsch, O., 315, 316.
Fritz, H., 302.
Fröhlich, H., 190, 328, 336.
Frost, R., 373.
Fukamoto, —, 261.

Gabriel, M., 30.
Gale, A. J., 225.
Galt, J. K., 225, 233, 234, 237.

* Not referred to in the text.

Garlick, G. F. J., 148.
Garten, W., 237.
Gärtner, H., 307, 319, 321.
Gaunt, J. A., 122, 140.
van Geel, W. Ch., 312, 314, 315, 326.
Gehrts, A., 301, 306, 318.
George, D. E., 428.
Gerlach, W., 80, 108.
Germer, L. H., 308, 322.
Ghose, B. N., 421.
Ghosh, B. B., 225.
Giacomini, A., 225.
Gibbons, J. J., 116, 131.
Gibson, A. F., 148.
Gilbert, J. H., 386.
Gill, E. W. B., 309.
Gilliland, T. R., 421, 423.
Ginsberg, N., 191.
Ginsburg, V. L., 227.
Gisolf, J. H., 317.
Glocker, R., 293.
Goeder, F. P., 233.
Golay, M. J. E., 165, 172.
Goldberg, L., 130.
Golden, S., 200.
Goldhaber, M., 22.
Goldman, I. M., 8.
Goldschmidt, H. J., 106.
Gollmick, H. J., 237.
Gombás, P., 131.
Good, W. E., 183, 185, 196–198.
Good, W. M., 28, 29.
Goodall, W. M., 415, 416, 418, 421, 423, 426.
Gordy, W., 186, 198, 199.
Gore, R. C., 174, 175.
Gorman, J. G., 200.
Gorodetzky, A., 226.
Gorokhovski, Y. N., 252, 265.
Goudet, G., 237.
Gourevitch, S. B., 225.
Green, A. L., 415.
Green, C. B., 163.
Greene, J. B., 75.
Gregg, E. C., 237.
Griffiths, E., 82, 377.
Grimley, T. B., 261.
Groenewold, H. J., 210.
Gronwall, T. H., 132.
Gross, E., 224.
Grube, G., 101.
Gruetzmacher, M., 158.
Gudden, B., 154, 158, 313.
Guében, G., 22.
Gugelot, P. C., 28.
Guggenheim, E. A., 109.
Gundermann, J., 313, 315, 316.
Gurevich, L., 232.
Gurney, R. W., 248, 262, 268, 270, 274, 276, 287, 289, 313–315, 317, 319–321, 337.
Gustafsson, G., 96, 97, 101.
Gyulai, Z., 264.

Hadley, L. N., 196.
Hafstad, L. R., 1, 3, 419.
Hahn, O., 20, 21, 23, 32.
Hahnemann, H., 237.
Hainer, R. M., 193, 199.
Hall, C. E., 287.
Hall, L., 221, 228.
Halpern, J., 36, 37.
Hamaker, H. C., 306.
Hamburger, L., 271.
Hamer, R., 263.
Hamilton, R. A. 423.
Hansell, C. W., 442*.
Hanson, A. O., 15, 16.
Hanson, W. T., 256.
Harang, L., 426, 427, 444, 448, 449.
Harding, S. T., 380.
Hardy, H. C., 237.
Hardy, J. D., 174.
Harley, L. S., 340*.
Harris, L., 161.
Harrison, G. B., 260, 261, 265.
Harting, H., 129.
Hartman, W., 315, 316.
Hartmann, —, 312, 317.
Hartree, D. R., 113, 118, 120–125, 127–131, 133, 138–140, 143.
Hartree, W., 120–123, 125, 127–129, 133, 138, 139, 143.
Hartung, E. J., 256.
Hassé, H. R., 132, 133.
Hauffe, K., 315.
Haurwitz, E. S., 123, 130.
Hautot, A., 280.
Hayes, H. V., 165.
Hearmon, R. F. S., 233.
Hebb, M. H., 35, 37.
Heintz, E., 166.
Heinze, W., 302–305, 309, 325, 329.
Heising, R. A., 416.
Helliwell, R. A., 441, 442.
Henderson, W. J., 21.
Hendrick, J., 386.
Henrich, L. R., 140.
Herb, R. G., 1, 3, 7, 15.
Herbert, L., 331.
Herlofson, N., 418, 422, 436–438, 440, 453.
Hermann, G., 300, 303, 306, 309, 319, 325, 329, 330.
Herscher, L. W., 174.
Hershberger, W. D., 185, 198, 199.
Herzberg, G., 411.
Herzfeld, K. F., 207, 209, 211, 213, 221, 225, 238, 244, 266.
Hess, B., 293.
Hess, V. F., 23.
von Hevesy, G., 312, 315, 320.
Hewlett, C. W., 151.
Hey, J. S., 401, 421, 428–434, 438, 440, 442.
Hiedemann, E., 205.
Hildebrand, E., 101.
Hill, A. G., 200.
Hillger, R. E., 183, 196, 200.
Hilsch, R., 257, 270.
Hirone, T., 89.
Hirschfelder, J. O., 218.
Hirschlaff, E., 291.
Hodgson, B., 340*.
Hodgson, M. B., 258.
Hoffleit, D., 401, 449.
Hoffmann, P. O., 276.
Hoffmeister, C., 397, 409, 444.
von Höhnel, F., 386.
Holden, A. N., 199, 200.
Holst, G., 149.
Holstein, T., 78.
Holtsmark, J., 129.
Holzman, B., 373, 377, 384.
Hopwood, F. L., 237.
Horn, F. H., 164.
Hornig, D. F., 161, 162.
Horton, F., 310.
Horton, R. E., 386.
Hoselitz, K., 104, 105.
Housekeeper, W. G., 14.
Houston, R. A., 237.
Houwink, R., 232.
Howard, L., 367.
Howell, A. H., 8.
Hsu, E. T-Y., 225.
Hubbard, D., 261, 273, 274.
Hubbard, J. C., 215.
Huber, H., 302, 303, 307, 309, 319, 321, 328, 331.
Huber, O., 28.
Hudson, C. M., 3.
Hudson, J. H., 272, 273.
Huggins, M. L., 257.
Hughes, A. L., 131.
Hughes, R. H., 186.
Hull, A. W., 307.
Hunter, J. L., 225.
Hunter, L. P., 233.
Huntington, H. B., 233, 237.
Hurdle, B. G., 236.
Huse, E., 265.
Hush, J. M., 16.
Huxford, W. S., 308, 311, 328.
Hyde, W. L., 163, 165.
Hylleraas, E. A., 131, 132.

Illsley, P. F., 100.
Ingham, A. C., 131.
Inglis, D. R., 143.
Ingram, L. J., 420, 423, 424, 426.
Isagawa, Y., 421, 423.
Isakovich, M. A., 223.
Isensee, H., 302, 312, 321.
Ito, Y., 418.
van Itterbeek, A., 215–218.

Jacobi, R. B., 414.
Jacobs, H., 306.
Jacobsen, J. C., 21, 39.
James, H. M., 116, 117, 129, 131, 132.
James, T. H., 272, 284–286, 288, 289.
Jatkar, S. K. M., 215.
Jauch, J. M., 198.
Jauncey, G. E. M., 279.
Jausseran, C., 252, 253.
Jeans, J. H., 208.
Jeffreys, B., 128.
Jeffreys, H., 369, 371, 372.
Jen, C. K., 198.

* Not referred to in the text.

Jetter, U., 233.
Joffe, A. F., 315, 317, 331.
Joffe, A. V., 315–317, 331.
Joffe, J., 317.
Johnson, E. A., 161, 162.
Johnson, J. B., 336.
Johnson, W. W. A., 413.
Joliot, F., 7, 21.
Jones, E. M., 225, 234.
Jones, H., 82, 109.
Jones, L. A., 251.
Jones, L. F., 427.
Jones, P. L. F., 225.
Jost, W., 259, 260.
Judson, E. B., 419, 421, 423, 427.
Junkelmann, R., 450.
Jurriens, H. J., 255.
Jusé, W. P., 313, 315, 316.

Kaiser, H. F., 171.
Kallmann, H., 217.
Kameyana, N., 261.
Kan-Chang Wang, 38.
Kantrowitz, A., 213.
Karlik, B., 23, 25.
Katchinsky, N. A., 368.
Kaufmann, A. R., 87, 96–100.
Kawamura, H., 304, 310, 311, 315.
Kê, T.-S., 232.
Keen, B. A., 384.
Keesom, W. H., 82, 93.
Keller, H., 215, 216.
Kempf, A., 277.
Kendall, H. C., 215.
Kennard, E. H., 122, 139.
Kerst, D. W., 1, 3.
Kessler, M., 186, 198.
Khastgir, S. R., 441.
Kieser, H., 271.
Kiessig, H., 293.
Kimball, G. E., 127.
Kincaid, J. F., 218, 219, 221.
King, A. S., 413.
King, F. H., 368.
King, G. W., 193, 199.
Kirby, S. S., 419, 421, 423, 427.
Kirchhoff, G., 207.
Kirkpatrick, H. A., 131.
Kittel, C., 205, 218, 220.
Kittredge, J., 367, 386.
Kleinschrod, F. J., 266, 267, 269.
Klemm, W., 85, 109.
Klöpf, A., 238.
von Klüber, H., 414.
Klug, H., 257.
Kneser, H. O., 215, 216, 221.
Kniepkamp, H., 303, 306.
Knight, H. de B., 331.
Knudsen, V. O., 217.
Koch, E., 259, 260.
Koch, P. P., 256, 293.
Kofoed-Hansen, O., 39.
Kögelmann, F., 277.
Kohler, M., 216.
Kohnke, H., 367.
Koller, L. R., 301.
Kolomiez, B., 152.
Kolthoff, I. M., 261.
Konomi, M., 421, 423.
Konopinski, E. J., 37, 41.
Kónya, A., 131.
Koopmans, T., 140.
Kornfeld, G., 251, 272, 277, 288.
Kornfeld, M., 237.
Koshal, R. S., 383.
Köster, W., 108.
Kovarik, A. F., 20.
Kozma, B., 131.
Kramers, H. A., 67, 115.
Krashnooshkin, P. E., 237.
Krebs, A., 23.
van Krefeld, A., 255, 268.
Kremers, H. C., 167, 168.
Kretchman, C. M., 254, 257.
Kroczek, J., 310, 311.
Kron, E., 252.
Kronig, R. de L., 140.
Krutter, H. M., 70.
Kuiper, G. P., 172.
Kurrelmeyer, B., 82, 93.
Kurtchatow, D., 22.
Kurtchatow, I., 22.
Kurtschatow, B. W., 313, 315, 316.
Kurtz, A. R., 237.
Kussmann, A., 80, 81, 94, 101.
Küster, A., 284.
Kyhl, R. L., 183, 191, 196, 197.

Labaw, L. W., 225, 237.
Lagemann, R. T., 225, 226.
Lainer, A., 255.
Lakshiminarayanan, D., 215.
Lamar, E. S., 13.
Lamb, W. E., Jr., 187.
Lambert, J. D., 215.
Landau, L., 213, 228, 232, 239.
Landau, R., 277.
Lane, C. T., 237, 240.
Lange, B., 158.
de Langhe, —, 276.
Langmuir, I., 308, 446.
von Laue, M., 299.
Lauwers, L., 215.
Lawes, J. B., 368, 386.
Lawson, A. W., 232.
Lazard, A., 7.
Leather, J. W., 378.
Le Blanc, M., 315.
Lecoin, M., 33.
Lecomte, J., 171.
Lee, C. H., 386.
Lee, E., 172, 198.
Leermakers, J. A., 259, 268.
Lehfeldt, W., 260–262, 291.
Leipunski, A. I., 35.
Lennard-Jones, J. E., 117.
Leontovic, M., 221.
Lescynski, W., 268, 277, 280.
Leslie, R. T., 310.
Levenson, G. I. P., 286, 287.
Lewis, W. B., 20–22, 162.
Lewis, W. V., 386.
Liddell, U., 175.
Liebold, W., 318.
Liepmann, H. W., 218.
Lifshitz, E., 239.
Lindemann, F. A., 444.
Lindsay, R. B., 123.
Linford, L. B., 308, 309.
Lippincott, J. B., 379.
Lipson, H., 84.
Liston, M. D., 174.
Livesey, D. L., 293.
Livingston, G. L., 378.
Lloyd, D., 367, 386.
Löhle, F., 257.
London, F., 217.
Loomis, G. A., 238.
Loosjes, R., 334.
Lorentz, H. A., 179, 188.
Loubser, J. H. N., 183, 186, 199, 201.
Loveland, R. P., 274.
Lovell, A. C. B., 408, 428–431, 434, 436–440, 452, 453.
Lowry, E. F., 301, 302, 304, 307, 318, 322, 340*.
Lübke, E., 310, 311.
Lucas, R., 223, 227.
Ludloff, H., 67.
Ludwig, J. B., 310, 336.
Lugeon, J., 378.
Lüppo-Cramer, H., 255, 277, 278.
Lyashenko, V. I., 317.
Lyman, E. M., 33.
Lynn, J. G., 238.

McDougall, J., 71, 123, 124.
Macfarlane, G. G., 158.
McGowan, F. K., 22.
McGrath, J. W., 237.
McIntosh, L. R., 1, 8, 13.
McIntosh, R. A., 390.
McKellar, A., 411, 412.
MacKenzie, K. R., 25.
McKibben, J. L., 15, 16.
McLauchlan, J. C., 125.
McSkimin, H. J., 228, 230, 231, 234.
Maddock, A. J., 302–305, 321.
Maeda, K., 421, 423.
Majorana, Q., 151.
Makarov, G., 225.
Maltzev, V. A., 394.
Mandelstam, L., 221.
Manders, C., 99.
Mandeville, C. E., 16.
Mann, M. M., 131.
Manning, L. A., 441, 442.
Manning, M. F., 75, 90, 122, 125, 127, 129.
Marian, V., 87, 97.
Mariëns, P., 216–218.
Maris, H. B., 418, 422, 438, 446.
Mark, H., 226.
Marriage, A., 254, 277–280, 291.
Marsden, E., 19, 20.
Martyn, D. F., 419.
Maskell, E. J., 369.
Maslach, G., 225, 234.

* Not referred to in the text.

Mason, W. P., 217, 222–224, 228, 230–234.
Massey, H. S. W., 444, 448, 450.
Matta, K., 215.
Maurer, R. J., 316.
Maurer, W., 450.
Maxwell, J. C., 223.
Mayer, J. E., 259.
Medicus, H., 28.
Mees, C. E. K., 259, 267, 268, 281.
Mees, R., 237.
Megaw, E. C. S., 335.
Meidinger, W., 269, 279, 284.
Meiklyar, P., 277.
Meitner, L., 32, 33.
Mendelssohn, K., 291.
Mercer, K. A., 428.
Mercier, P., 20.
Merigoux, R., 256.
Merrill, G. P., 414.
Merriman, T., 386.
Merritt, F. R., 193, 199, 200.
Merton, G., 402.
Meulendyke, C. E., 265.
Meyer, A. F., 378.
Meyer, S., 23.
Meyer, W., 310, 311, 315, 316, 321.
Mikhailov, I. G., 225.
Milatz, J. M. W., 162.
Millard, —, 127.
Miller, A. E., 238.
Miller, N. F., 147.
Millington, G., 427.
Millis, W. T., 340*.
Millman, J., 125.
Millman, P. M., 401, 412, 413.
Milton, R. M., 164.
Mimno, H. R., 415, 421, 427.
Mindlina, A., 225.
Minkowski, R., 411.
Minohara, T., 418.
Mitra, S. K., 421.
Mizuta, K., 261.
Mögel, H., 421.
Möglich, F., 316.
Mokhtar, M., 215.
Møller, C., 35, 37.
Mollet, P., 166.
Mönch, G., 317.
Moon, P., 164.
Moore, G. E., 252, 277, 341*.
Moore, H. R., 163.
Morgulis, N., 329, 335.
Morrison, J., 302, 315, 329.
Morse, P. M., 123, 130, 131.
Morton, G. A., 149.
Moss, T. S., 154.
Mott, N. F., 69, 82, 89, 109, 248, 261, 262, 268–270, 274, 276, 287, 289, 312–315, 317, 319–321, 334, 337.
Mouzon, J. C., 450, 451.
Mueller, D. W., 7, 308.
Murgoci, R., 301, 307, 310, 320, 322, 334.
Murray, J., 386.
Mutter, E., 271.
Mutter, W. E., 334.
Myers, F. E., 15.
Myers, L. M., 149.
Myssowsky, L., 22.

Nafe, J. E., 279.
Nagaoka, H., 417.
Nagorsky, A., 335.
Naismith, R., 416, 419, 420, 422–424, 426–429, 431, 438, 441, 442.
Natanson, L., 223.
Natterer, K., 7.
Nebel, C., 303, 366.
Néel, L., 68, 69, 74, 76, 77, 80, 83, 87, 91, 97, 99, 100, 109.
Neldel, H., 315, 316.
Nelson, C. N., 251.
Nielsen, H. H., 197.
Nielsen, J. R., 175.
Niessen, K. F., 89.
von Niessl, G., 393, 397.
Nishibori, E., 304, 310, 311, 315.
Nix, F. C., 152, 153.
Noddack, I., 413.
Noddack, W., 259, 271, 413.
Nolle, A. W., 233.
Norman, D. P., 413.
Norrish, R. G. W., 254.
Noury, J., 225.
Nurmi, U., 233.
Nutting, P. G., 284.

O'Brien, A. S., 261.
Occhialini, G. P. S., 293.
Öpik, E., 392, 394, 397, 399, 446, 447, 450, 451.
Oertel, L., 67.
O'Keefe, B. J., 161, 162.
Olivier, C. P., 390, 391, 393, 396–398, 402, 409.
Olmsted, D., 393.
Opechowski, W., 67.
Oppenheimer, J. R., 21.
Orthmann, W., 33.
Osida, I., 223.
Ostwald, W., 282.
Ott, —, 264.
Overbeck, C. J., 215.
Owen, E. A., 106.
Oxley, C. L., 154.

Pan, S. T., 99, 100.
Paneth, F., 23.
Paneth, F. A., 414.
Parker, R. C., 172.
Parkinson, D. B., 1, 3, 15.
Parshad, R., 225, 237.
Parsons, S. T., 431, 434, 442.
Päsler, M., 229.
Pasquil, F., 372, 373, 377.
Paul, F. W., 172.
Pawsey, J. L., 422.
Peacock, W. C., 28, 29.
Pearce, R. R., 83, 86, 100.
Pearson, G. L., 163, 164.
Peaslee, D., 28, 29.
Peierls, R., 42.
Pekar, S. I., 317.
Pekeris, C. L., 231.
Pelc, S. R., 293.
Pellam, J. R., 225, 234, 236, 237, 240.
Pellett, H. E., 238.
Penman, H. L., 375, 381–383.
Penrose, R. P., 181, 183, 185, 186, 190, 196, 199–202.
Pereira, H. C., 368.
Perey, M., 22.
Perkins, P. B., 20.
Peshkov, V., 239.
Petersen, H., 140.
Petersen, O., 237.
Peterson, H. O., 442*.
Petrashen, M. J., 117.
Pfriem, H., 213.
Pfund, A. H., 160, 168.
Phillips, F. S., 163.
Phipps, T. E., 310.
Pickard, G. W., 418.
Pickles, A. T., 105.
Piddington, J. H., 424, 426, 438.
Pielemeier, W. H., 215, 237.
Pierce, J. A., 421, 422, 426, 427, 431, 436, 438, 441, 450, 452.
Pinoir, R., 266.
Plendl, H., 421.
Podszus, E., 315.
Pohl, R. W., 167, 257, 262, 270, 291.
Poisson, S. D., 223.
Polder, D., 197.
Pomeranchuk, J., 232.
Pomerantz, M. A., 333, 335.
Pond, T. A., 202.
Pontecorvo, B., 42.
Porter, A., 127.
Porter, J. G., 398–400, 438.
Potter, H. H., 80, 81.
Poulsen, G. G., 6.
Powell, C. F., 293.
Powell, R. W., 369, 377, 378.
Prasad, R., 225.
Pratt, O. S., 340*.
Pratt, T. H., 149.
Preiswerk, P., 28.
Prentice, J. P. M., 390, 394, 428–430, 434, 437, 438, 448, 452, 453.
Prescott, C. H., Jr., 302, 315, 329.
Prescott, J. A., 377.
Primakoff, H., 78, 209.
Pulley, O. O., 419.
Pumper, E. J., 216.

Quäck, E., 418.
Quigley, T. H., 215.
Quimby, S. L., 233.
Quinn, J., 225.
Quirk, R. F., 252, 268.

Rado, G. T., 87.
Ragona, D., 367.
Rama Rao, M., 225.
Ramachandra Rao, B., 237.
Raman, C. V., 224.
Raman, P. K., 381.
Ramberg, E., 122, 139.
Ramm, W., 224.

* Not referred to in the text.

Randall, H. M., 191.
Randall, J. T., 148, 331, 335.
Rank, D. H., 224.
Ranzi, I., 419.
Rapuano, R. A., 225, 227, 234, 236.
Ratcliffe, J. A., 415, 419, 420, 422, 423, 427.
Rawling, S. O., 254, 273.
Read, T. A., 232.
Reiger, R., 222.
Reimann, A. L., 299–301, 305–312, 315, 317–320, 322–324, 334.
Reinders, W., 256, 271, 281, 282.
Rendall, G. R., 225.
Renwick, F. F., 268.
Reymer, S. E., 421.
Reynolds, N. B., 308.
Rhodes, J. E., Jr., 213.
Rice, F. O., 207, 211, 218.
Richards, W. T., 212.
Richardson, B., 376.
Richardson, E. G., 215, 238, 239.
Richardson, O. W., 299, 308.
Rieckmann, P., 226.
Ring, H., 199.
Ringo, G. R., 226, 236.
Ritchie, M., 256.
Roberts, A., 200.
Robertson, H. P., 407.
Rock, G. D., 225, 227, 238.
Roder, H., 166.
Rodine, M. T., 7.
Roess, L. C., 161, 162, 174.
Rohwer, C., 367, 377–380.
Rona, E., 22.
Rooksby, H. P., 304, 318, 333.
Rose, A., 308.
Rose, M. E., 213.
Rosenblum, S., 20.
Ross, F. E., 254, 255.
Rothe, H., 301, 312.
Rothenbach, M., 23.
Rousseau, E., 315.
Roussinow, L., 22.
Rowlands, S., 30.
Roy, A. S., 213.
Rua, E., 252, 253.
Rüfer, W., 225.
Rumer, G., 228, 232.
Russell, H. N., 405.
Russell, J. V., 238.
Rutgers, A. J., 225, 238.
Rutherford, Lord, 18*, 20–22, 32.

Sabattier, 254.
Sachse, H., 315.
Safford, F. J., 7, 15.
Sahni, P. N., 382.
Sakadi, Z., 226.
Saneyosi, Z., 213.
Satakopan, V., 381.
Savostjanova, M., 272.
Saxton, J. A., 193.
Schade, O. 331, 332.
Schaeberle, J. M., 398.
Schafer, J. P., 415, 416, 418, 421, 423, 426.
Schamberg, R., 209.
Schaum, K., 293.
Scheffer, F. E. C., 287.
Scheibe, G., 269.
Scherrer, P., 26, 28.
Schiffer, M., 225.
Schilling, H. K., 215.
Schintlmeister, J., 22.
Schlapp, R., 193.
Schmid, G., 226.
Schmidt, A., 310, 311, 321.
Schmidt, E., 233.
Schoen, A. L., 287.
Schoen, E., 81.
Schofield, R. K., 381, 385.
Scholp, A. C., 161.
Schonne, P., 266.
Schopper, E., 293.
Schottky, W., 158, 299, 308, 312, 317, 319, 325.
Schriel, M., 311.
Schröpel, H., 315.
Schubotz, F., 238.
Schulman (Shul'man), A. R., 315, 341.
Schultz, H., 240.
Schulze, A., 80.
Schumb, W. C., 8.
Schwarz, G., 256.
Seaborg, G. T., 24, 26.
Sears, R. W., 310, 321, 322.
Segrè, E., 23, 25.
Seith, W., 320.
Seitz, F., 109, 230, 316.
Selwood, P. W., 109.
Selwyn, E. W. H., 248.
Shantz, H. L., 368.
Shapiro, I., 261.
Shapley, H., 397.
Shaw, W. N., 369.
Shen, S. T., 130.
Sheng, H.-Y., 196.
Shenk, J. H., 169.
Shenstone, A. G., 132.
Sheppard, S. E., 252, 267, 268, 270–274, 281, 283.
Shive, J. N., 165.
Shortley, G. H., 113, 115, 117, 120, 122, 127–129, 143.
Shoup, W. E., 116.
Shrader, E. F., 199.
Shul'man (Schulman) A. R., 315, 341.
Siegel, S., 233, 234.
Siegel, —, 27.
Silberstein, L., 271, 272.
Simmons, J. W., 199.
Simon, E. N., 238.
Simon, H., 318.
Simpson, O., 159, 160.
Singh, B. K., 225.
Sivian, L. J., 212.
Skellett, A. M., 417, 418, 421, 423, 424, 426, 450.
Skinner, H. W. B., 93.
Slater, J. C., 68, 70, 115, 118, 120, 122, 131, 133.
Slater, 236.
Sleight, R. B., 379, 384.
Smekal, A., 233, 310.
Smith, A. G., 199.
Smith, D. C., 175.
Smith, F. W., Jr., 237.
Smith, M. C., 225.
Smith, N., 421.
Smith, R. A., 450.
Smith, W. V., 199, 202.
Soddy, F., 19–21.
Sokolov, —, 238.
Sollner, K., 238.
Sommer, A., 149.
Soole, B. W., 160.
de Sorbo, W., 164.
Sosnowski, L., 154, 159, 160.
Southwell, R. V., 127.
Spanner, H. J., 310–312.
Sparrow, C. M., 444, 445.
Spencer, L. van C., 293.
Spenke, E., 312, 317.
Sperduto, A., 1, 8, 13.
Sproull, R. L., 305, 309, 331.
Squire, C. F., 240.
Starkiewicz, J., 151, 159, 160.
Starr, C., 96–98, 100.
Stasiw, O., 257, 262, 263, 266.
Staud, C. J., 259, 268.
Stefan, J., 369.
Steiner, W., 266.
Steinhardt, L. R., 164.
Steinhaus, W., 81.
Sterry, J., 254.
Stevens, G. W. W., 254, 263, 277–281, 291.
Stevenson, A. F., 114, 132.
Stevenson, D. P., 218.
Steward, E. G., 318.
Stewart, E. S., 214–216.
Stewart, J. L., 215, 237.
Stewart, J. S., 428–434, 438, 440, 442.
Stockbarger, D. C., 167.
Stoffregen, W., 427.
Stokes, G. G., 207.
Stoner, E. C., 43, 70, 71, 77, 82, 84, 109.
Strandberg, M. W. P., 183, 196, 197, 200.
Strock, L. W., 252, 276.
Strömgren, E., 402.
Strong, J., 161.
Stuart, D. M., 421, 427.
Stumpf, P. K., 237.
Sucksmith, W., 52, 83, 86, 100, 104–107.
Sully, A. H., 106.
Sunanda Bai, K., 224.
Suryan, G., 226.
Sutherland, G. B. B. M., 144, 148, 159, 160, 167, 172–175, 198.
Sutton, O. G., 372.
Sutton, R. M., 450, 451.
Svedberg, T., 258, 271.
Sverdrup, H. V., 373, 375, 376.
Swann, W. F. G., 237.
Sweeney, W. J., 174.
Swings, P., 411.
Swirles, B., 123, 128, 130, 139, 143.

* Not referred to in the text.

Sykes, C., 81, 82.
Syam, P., 419, 421, 423.
Symons, G. J., 366.
Szilard, L., 21.

Takahari, M., 315, 317, 322, 327.
Taylor, A., 94, 106.
Taylor, A. H., 427.
Taylor, A. M., 253.
Taylor, G. I., 371, 372.
Teeter, C. E., Jr., 225, 234, 237.
Telfair, D., 215, 216, 237.
Teller, E., 213, 269.
Teltow, J., 257, 262, 263, 266, 269.
Te-Tchao, O., 225, 227.
Thatcher, W. A., 139, 140.
Thiede, H., 233, 237.
Thielmann, K., 94.
Thom, J. A., 256.
Thompson, F. C., 30.
Thompson, H. W., 174, 175.
Thompson, N., 93.
Thornthwaite, C. W., 373, 384.
Thys, L., 216.
Timoshenko, S., 242.
Timovev, A. W., 335.
Tisza, L., 213, 221, 239, 240.
Todesco, G., 151.
Tolansky, S., 199.
Tollert, H., 268, 280.
Tonks, L., 218.
Torrance, C., 126.
Townes, C. H., 185, 186, 193, 196, 199, 200, 202.
Townsend, C. D., 238.
Toy, F. C., 260–262, 265, 274.
Traub, W., 293.
Treloar, L. R. G., 305, 310–312, 324.
Trent, H. M., 238.
Treptow, A. W., 152, 153.
Treu, M., 154, 158, 159.
Trillat, J. J., 256.
Trivelli, A. P. H., 250, 271, 272, 274, 280.
Trombe, F., 85, 86.
Trump, J. G., 1, 5, 7, 9, 15, 17.
Tsien, H., 209.
Tubandt, C., 261.
Turkevitch, J., 199.
Turner, C. M., 3.
Turner, L. A., 23, 25.
Tuve, M. A., 1, 3, 415, 419.

Uehara, Y., 315, 317, 322, 327, 328.
Uhlenbeck, G. E., 37, 41.
Undheim, B., 131, 132.
Ungar, G., 268.
Urbach, F., 147, 171, 256.
Urbain, G., 84.
Urquhart, R. C., 8.

Valentiner, S., 93.
Van Atta, L. C., 1, 7, 15.
Van de Graaff, R. J., 1, 5, 8, 9, 13, 15.
Vandoninck, W., 215.
Vane, A. B., 191.
Vanselow, W., 252.
Van Vleck, J. H., 53, 68, 128, 179, 188, 190, 193, 197, 198, 216.
Vassy, A., 266, 282.
Vassy, E., 266.
de Vaucouleur, G., 276.
Veenemans, C. F., 307, 319.
Vengerov, M., 238.
Venkataraman, K., 441.
Venkateswaran, C. S., 224.
Vermaelen, R., 215.
Villard, O. G., 441, 442.
Villars, D. S., 301.
Vink, H. J., 334.
Vinti, J. P., 131.
Viollet, R., 276.
Visentini, M., 367.
Vogler, H. J., 256.
Vogt, E., 101.
Vogt, W., 312.
Voigt, A. F., 23.
Volmer, M., 289.
van Voorhis, C. C., 446.
van Voorhis, S. N., 30.
Vorontsov-Velyaminov, B., 404.
Vsessviatsky, S., 404.

Wadia, H. M., 378.
van der Waerden, 120.
Wagener, S., 300, 302–307, 309, 319, 321, 325, 328–330.
Wagner, C., 259, 260, 313, 315, 316.
Waller, I., 118.
Walling, E., 21, 26.
Walter, J. E., 185, 198.
Walther, H., 231.
Wambacher, H., 293.
Wang, Kan-Chang, 38.
Wang, P. H., 421.
Wannier, G. H., 200.
Ward, F. A. B., 21.
Warren, R. E., 3.
Waterhouse, J., 254.
Watson, Fletcher, 392, 441.
Watson-Watt, R. A., 419, 424.
Watt, E. E., 16.
Watts, J. M., 428.
Watts, R. J., 186, 198.
Webb, J. H., 253, 268, 271, 272, 276, 277, 279, 289–292.
Weber, L. R., 191, 233.
Wegel, R. L., 231.
Wehnelt, A., 300.
Weigle, J., 233.
Weinland, C. E., 253.
Weiss, P., 84.
Weisskopf, V. F., 180, 188, 198.
von Weizsäcker, C. F., 22, 30.
van der Welden, H. A., 162.
Wells, H. W., 419, 421, 423, 428.
Wentink, T., Jr., 183, 196, 200.
Wertenstein, L., 37.
West, W., 262, 267.
West, W. A., 232.
Whatman, A. B., 423.
Wheatcroft, E. L. E., 309.
Wheeler, M. A., 97–99.
Whipple, F. L., 396, 409, 438, 444, 446, 447, 449.
White, E. L. C., 419, 421, 427.
White, J. W., 169.
White, W. N., 385.
Whitelaw, N. G., 128.
Whitney, B. S., 398.
Widell, E. G., 333.
Wightman, E. P., 252, 268, 272.
Wilkins, A. F., 419, 424.
Wilkins, M. H. F., 148.
Wilkinson, H., 81, 82.
Willard, G. W., 225, 237.
Williams, A. O., 130.
Williams, A. O., Jr., 225.
Williams, D., 186, 198.
Williams, E. J., 214.
Williams, J. D., 392.
Williams, N. H., 180, 195, 311.
Williams, V. Z., 175.
Williamson, R. E., 116, 139, 140.
Willis, F. H., 225.
Wills, L. A., 116.
Willshaw, W. E., 331–333.
Wilsey, R. B., 258.
Wilson, A. H., 109, 312–315
Wilson, C. T. R., 419.
Wilson, E. B., Jr., 186, 197, 200.
Wilson, R. H., 20.
Wilson, W., 172, 261.
Wilson, W. S., 123.
Winkler, O., 101.
Wolfe, R. N., 255.
Wollny, E., 368.
Wood, R. W., 253, 280.
Woodward, W. M., 13.
Wooster, W. A., 33.
Wright, D. A., 310, 319, 334.
Wright, N., 167, 174.
Wright, P., 21.
Wu, C.-K., 198.
Wu, T.-Y., 130.
Wul, B. M., 8.
Wüst, G., 378, 387.
Wynn-Williams, C. E., 20, 21.

Yost, W. J., 124, 127.
Young, A. F. A., 308.
Young, L. A., 123, 130.
Young, L. C., 427.

Zachoval, L., 224.
Zahl, H. A., 165.
Zener, C., 130, 213, 217, 228–230, 232.
Zenker, —, 253.
Zimens, K. E., 261.
Zlotowski, I., 33.
Zvetkov, V., 225.
Zwemer, R. L., 238.

PRINTED BY TAYLOR AND FRANCIS, LTD.,
RED LION COURT, FLEET STREET,
LONDON E.C. 4.

ELECTROSTATIC GENERATORS FOR THE ACCELERATION OF CHARGED PARTICLES

By R. J. Van de Graaff, J. G. Trump and W. W. Buechner,
Massachusetts Institute of Technology, Cambridge, Massachusetts

§1. INTRODUCTION

The need, much emphasized by Rutherford (1927), for copious sources of charged bodies similar in energy to the particles emitted from naturally radioactive materials remains an active requirement in nuclear physics and other fields. In general, it is desirable not only that such particles be available in adequate amounts and with sufficient energy to penetrate the atomic nucleus, but also that they be homogeneous and steady in energy and that they emerge from the apparatus in which they are produced in a parallel beam with little accompanying stray radiation. Moreover, it should be feasible to measure the voltage of such particles with great accuracy and to vary it readily over a wide range. In the energy range up to 10 million volts, which is sufficient for investigating many of the fundamental phenomena of nuclei throughout the periodic table, these rather rigorous requirements for precise nuclear research can be met with directness and simplicity by the combination of an electrostatic generator and an accelerating tube. Examples of such combinations are those described by Van de Graaff (1931), Van de Graaff, Compton and Van Atta (1933), Tuve, Hafstad and Dahl (1935), and Herb, Parkinson and Kerst (1937). For those nuclear investigations where energies of higher ranges are required, more indirect methods of acceleration must be employed, as in the case of the cyclotron, betatron and synchrotron.

The electrostatic generator and accelerating tube can likewise be employed with advantage in non-nuclear applications. For radiation therapy, penetrating roentgen rays characterized by high skin tolerance, and particularly effective for the irradiation of deep tumours, can be produced by compact generators operating at several million volts (Trump and Van de Graaff, 1939; Trump and Cloud, 1943). For industrial radiography, the intensity and penetrating nature of such radiation increases radiographic speed, definition, sensitivity, and latitude in the examination of heavy metal sections, as has been shown by Van de Graaff, Buechner, Feshbach, Sperduto, Burrill, and McIntosh (1946). For this radiographic application, the constant voltage makes possible the concentration of the electron beam on an exceedingly small target focal spot with subsequent point-source emission. Such steady controlled sources of charged particles amenable to accurate application and measurement are indicated for many physical, biological, and photochemical investigations.

§2. PRINCIPLES OF ELECTROSTATIC BELT GENERATORS

Early frictional and influence machines have been replaced for these purposes by a more direct and effective type of electrostatic generator which employs a belt as the charge-conveying means. The principle of this generator can be described

by making reference to the diagram shown in figure 1. The generator consists of a well-rounded high-voltage terminal supported from ground on an insulating column, and of a charge-conveying system consisting of one or more belts of insulating material running between this terminal and ground. Electric charge of one polarity is sprayed by corona on the belt at its grounded end and is transported by the belt into the hollow terminal, where it is removed. Within the terminal, charge of the opposite polarity may be deposited on the belt and transferred to ground. At any instant the terminal potential is $V = Q/C$, where Q is the stored charge and C the capacitance of the terminal to ground. The terminal

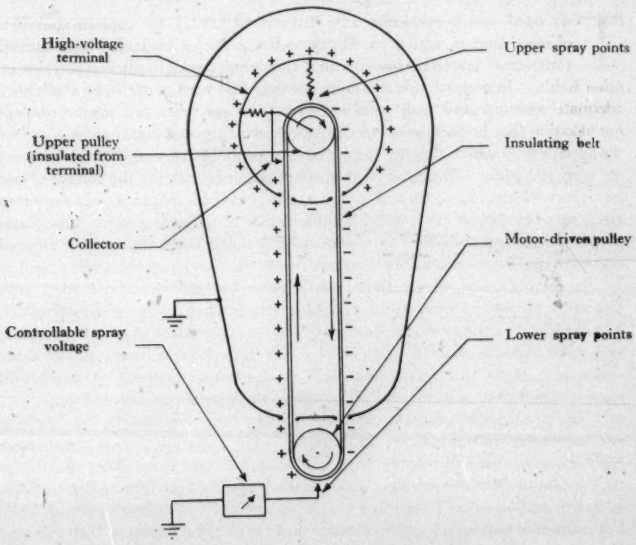


Figure 1. Schematic diagram of electrostatic belt generator.

potential rises initially at the rate $dV/dt = i/C$, where i is the net charging current to the terminal. This rate is commonly of the order of one million volts per second.

The terminal will maintain that equilibrium potential at which the charging current delivered by the belt is just balanced by the load current. The load current may consist of the accelerating tube current, current down the column which supports the terminal, as well as corona current from the terminal itself. These latter two components of load current increase with terminal voltage and ensure its stability around the equilibrium value. By adjusting the current carried by the belt and the load current, the terminal voltage may be established at any desired value up to the maximum. An essential feature of these generators is that the function of belt-charging and charge-removing is performed within the field-free

spaces existing below the ground plane and within the high-voltage terminal. In this way, these functions are made independent of the potential difference which may exist between the terminal and ground.

The maximum voltage which can be maintained on the terminal depends only on its geometry and the dielectric strength of the surrounding insulating medium. The terminal shape must be such that, at the desired operating voltage, the surface gradient at all points is well below the ionizing or breakdown value. Ideally considered, an isolated conducting sphere is the most favourable electrode shape and would maintain a voltage Er, where E is the permissible gradient for the insulating medium surrounding the sphere and r its radius. The necessity, however, of physically supporting the high-voltage terminal from ground, as well as of terminating an electrically charged belt and an acceleration tube inside the terminal, makes it desirable to modify its spherical shape in the region of the column. These features are shown in figure 2, which is a photograph of an early electrostatic accelerator, built and used very successfully for nuclear research at the Carnegie Institution of Washington (Tuve, Hafstad and Dahl, 1935).

Likewise, the supporting column, which ideally might be a solid insulating rod, must be modified because of the field-distorting influence of the charged belt. This is particularly true in compressed-gas insulated generators if maximum advantage of the improved insulating medium is to be obtained. It is best accomplished by constructing the column of a series of spaced conducting members or planes separated by insulating supports. Corona or resistances between the equipotential planes are used to divide the terminal potential uniformly along this column. In this way the electric field in the region of the belt and tube and in the vicinity of the high-voltage terminal and the ground plane is maintained substantially uniform. The multiple acceleration tube may, with benefit, be mounted within this field-controlled column, its insulating length corresponding with the column, the ion or electron source being within the terminal, and the target electrode beginning at the ground plane. The total terminal potential is thus divided into a series of equal low-voltage increments and is applied in a uniform manner to both belt and multi-acceleration tube. Figure 3 shows a practical embodiment of this terminal and column arrangement. Both vertical and horizontal column constructions, the latter based on the developments at the University of Wisconsin (Herb, Parkinson and Kerst, 1937) and illustrated in figures 4 and 5, are now commonly employed.

The maximum terminal voltage is influenced by the terminal and tank geometries. For the concentric spherical arrangement ordinarily found in compressed-gas insulated generators, it can be shown that the maximum voltage is attained when the tank diameter is twice that of the high-voltage terminal. This assumes that breakdown ensues when a given terminal gradient is reached. In higher voltage generators it may become desirable to control the electric field around the terminal by means of intermediate metallic equipotential shields (Herb, Turner, Hudson and Warren, 1940). These derive their potential from the appropriate sections of the column to which they are connected, as illustrated in figures 4 and 11. Assuming again that the flashover of a gaseous gap is determined by the maximum electrode gradient, it is evident that the spacings and potentials of these intermediate shields should be such as to result in the same maximum

gradient values for each of the several gaps. This is realized if the radii of the electrodes, considered as concentric cylinders, obey the relation

$$\frac{r_1}{r_2} = \frac{r_2}{r_3} = \frac{r_{n-1}}{r_n},$$

where $r_1, r_2, \ldots, r_n$ represent the radii of the electrode surfaces, beginning at the terminal. Actually, the shields are combinations of cylindrical and spherical surfaces, but the cylindrical sections play the dominant rôle. In the limit, the effect of such shields is to render the gradient uniform throughout the gap from terminal to tank and to remove any restriction on the size of the terminal relative to the tank. In practice, however, it has not yet been desirable to use more than two intermediate shields.

§3. THE CHARGING SYSTEM

(i) *Belt materials*

Flat endless belts of insulating material have thus far proved the most effective charge-conveyors for both air and compressed-gas insulated electrostatic generators. Such belts are usually made of multi-ply rubber fabric, though woven cotton, nylon, and other materials possessing good mechanical and electrical properties have been used. The belt travels between the ground plane and the interior of the high-voltage terminal at constant speed, commonly between 3000 and 8000 feet per minute.

(ii) *Belt-charging methods*

Referring again to figure 1, electric charge is sprayed on the belts at the grounded end from a row of corona points extended across the width of the belt and directed at the pulley. A corona spray voltage maintained between these points and the pulley results in gaseous ionization at the points and causes a transfer of electric charge toward the pulley. This charge is deposited upon the intervening insulating belt, neutralizing any residual charge of opposite polarity remaining from the downward belt run, and leaving a net charge which is conveyed to the terminal. The charge sprayed on the belt in this manner is controlled by regulating the voltage between the points and the pulley and has a maximum value which depends on the insulating strength of the gaseous medium in which the belt operates. The belt charge may be removed within the terminal from the ascending belt by a corona-point collector connected to the terminal. In this arrangement only the upward belt run is utilized. The generated electric power is equivalent to the mechanical power required to move the belt charge to the terminal potential.

Within the high-voltage terminal a second source of corona-spray voltage may be used to neutralize the charge brought up and to charge the descending run of belt with opposite polarity. A simple self-inducing arrangement may also be used to perform this function. This is accomplished by insulating the upper pulley and maintaining it at a potential higher than that of the terminal by receiving the charge from the collector connected to the pulley as in figure 1. It is evident that on entering the hollow conducting terminal, the belt charge acquires a voltage relative to the terminal which is independent of the potential of the terminal with respect to ground. This enables a corona-point collector mounted within the

terminal and just below the point of tangency of the arriving belt to remove through ionization nearly all of the charge on the arriving belt. A second row of corona-spray points directed at the belt-covered upper pulley utilizes this voltage to ionize the gas at the points and effect a continuous transfer of electric charge from the terminal to the moving belt. This upper spray neutralizes any residual belt charge and leaves an excess of charge of the opposite polarity to be carried away. The arrangement is completed by placing a resistance between the upper pulley and the terminal, so adjusted in value that the current carried on the downward run of belt corresponds approximately to the current arriving on the ascending run. Thus both belt runs can be effective in transferring charge continuously between ground and high-voltage terminal.

(iii) *Permissible belt-charge density*

The maximum charge density which can be carried by a belt in air at atmospheric pressure is that which will result in a gradient of about 30 kilovolts per centimetre. If the electric field due to belt charge is directed normally in both directions from the belt surface, the maximum charge density is $5{\cdot}30 \times 10^{-9}$ coulombs per square centimetre. In practice, 40 to 60 per cent of this maximum value is commonly attained in generators which operate in air at atmospheric pressure.

While the most effective means of increasing the current capacity of an electrostatic generator is to increase the dielectric strength of the insulating medium in which it operates, common methods for securing greater current capacity in air-insulated generators are to increase the speed, width, and number of belts. In the operation of multiple belts, it is necessary that adjacent belt runs move in opposite directions and carry charge of opposite sign. The gradient at the surface of each belt due to its charge can be calculated according to Gauss's law and cannot exceed the value at which breakdown occurs. Evidently in a multiple-belt system, if each negatively charged belt has as its neighbour on both sides a positively charged belt, the gradient due to the surface charge will exist on both sides of the belt and will extend only to its nearest neighbour. A simple consideration of Gauss's law shows that the net current capacity of several adjacent belts carrying charge of the same sign cannot exceed that of a single belt unless each belt is isolated by conductors as discussed below.

(iv) *Control of the electric field due to belt charge*

If the charge distribution is assumed to be uniform on all belt surfaces, the outward-directed gradient on both sides of the belt in air at atmospheric pressure may be over 10,000 volts per centimetre. With superior insulating media this surface gradient may be proportionately higher; in 27 atmospheres of nitrogen, gradients normal to the belt surface as high as 100,000 volts per centimetre may be maintained. The transverse gradient due to belt charge is rarely a source of difficulty at atmospheric pressure, but may be the cause of serious flashover difficulties and current-carrying limitations at higher pressures unless suitable control is established of the fields due to belt charge.

This control can be accomplished by means of conducting rods connected to the column planes. Such an arrangement has been described by Trump and Van de Graaff (1939) and has the effect of localizing the electric field due to the

belt charge and of confining it to a region closely adjacent to the belt surfaces. These field-control rods are closely spaced along the column parallel to each face of the belt and are connected either directly or through a resistance to the equipotential plane of the column at that level. Figure 6 shows the electric-field map for an electrically charged belt within a controlled column (Poulsen, 1940).

Evidently the effect of these control rods is to reduce to a relatively small value the variation in potential in the transverse plane due to belt charge. With a gradient due to this charge of 100 kilovolts per centimetre, a field-control rod placed at 1 cm. from the belt in that plane would limit the potential difference to 100 kilovolts, whereas with a separation of 10 cm. between the belt surface and the equipotential plane at any given level it would tend to create a potential difference of

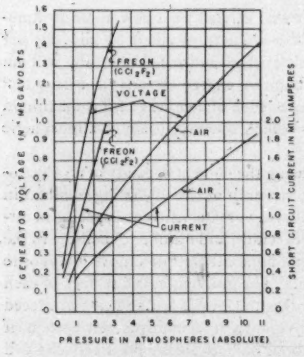


Figure 7. Curves showing the increase of maximum terminal voltage and of short-circuit current with gas pressure of air and of freon. The medical X-ray generator shown in figure 3 was used in these studies.

(*Reproduced, by kind permission, from the 'Physical Review'.*)

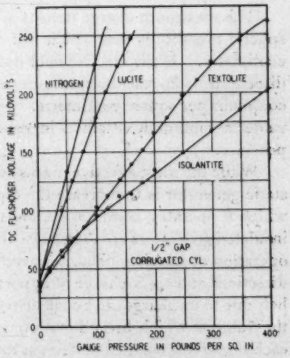


Figure 9. Direct-current flashover strength of nitrogen and of corrugated cylinders of lucite, textolite, and isolantite immersed in nitrogen at high pressures. Uniform field, ½-in. gap.

(*Reproduced, by kind permission, from the 'Transactions of the American Institute of Electrical Engineers'.*)

1 million volts. This arrangement of field-control rods closely spaced to each side of each run of belt confines the electric field due to belt charge and at the same time has the effect of dividing the belt lengthwise into a series of short sections with controlled potential applied to each. With such control, a fairly proportionate increase with gas pressure of current capacity and breakdown strength lengthwise along the belt may be secured, as illustrated in figure 7. Since the length of the generator column is governed to a large extent by the length of belt required to insulate the desired terminal voltage, this increase in dielectric strength along the belt is significant. Moreover, the requirement of equal charge on each run of belt, which is difficult to realize for all currents and all pressures, is eliminated by the shielding effect of the conducting rods which pass between the two runs of belt.

When the belt is carrying large amounts of charge, there are very considerable mechanical forces tending to draw it to the adjacent control rods. In order to minimize these forces and to prevent the rubbing of the belt on these conductors, it is necessary to maintain the belt in the electrical and mechanical neutral-plane midway between the control rods. This may be accomplished by surrounding occasional conducting rods with an insulating tube. Alternatively, separate insulating rods of glass, porcelain, or solid plastic material may be provided periodically along the belt travel.

§4. INSULATORS FOR ELECTROSTATIC GENERATORS

(i) *Gaseous insulating media*

Undoubtedly the most effective means for increasing the voltage and current capacity of an electrostatic accelerator is to increase the dielectric strength of the medium in which it operates. It was stated by Paschen, and it followed from Townsend's and subsequent analyses, that for a given gas and a uniform field the sparking voltage is a function of the product of the inter-electrode gap and the gas pressure. The more general similarity law declares that the sparking potential of a gas remains unchanged if the physical dimensions of the electrodes and gap are changed in the inverse ratio to the change in pressure. General realization that these laws of insulating strength versus pressure held fairly accurately to pressures of at least ten atmospheres, and departed only slowly thereafter, led to the early use of compressed gases such as air, nitrogen, and carbon dioxide for the insulation of electrostatic generators intended for operation in the higher voltage and current ranges (Barton, Mueller and Van Atta, 1932). Of these gases, air has the advantage of availability but the disadvantages of fire hazard, particularly at high pressures, and of the gradual corrosive influence of the compounds resulting from ionization. Carbon dioxide is subject to some question because of its possible decomposition into carbon monoxide and oxygen. Nitrogen is readily available, is closely similar in dielectric performance to air, eliminates fire hazard, and makes possible several times longer belt life because of absence of decomposition products.

That certain compounds in the gaseous state—notably those which contain chlorine, fluorine, or other negative atoms—have a higher insulating strength than air has been known since the investigations of Natterer (1889). The vapour pressure of carbon tetrachloride (CCl_4), which at room temperature is but 2 lb./sq. in., has long been known to increase markedly the dielectric strength of air at atmospheric pressure, and this increase is maintained at approximately the same value when the air is raised to higher pressures (Joliot, Feldenkrais and Lazard, 1936; Rodine and Herb, 1937). Charlton and Cooper (1937) have investigated the insulating properties of a number of similar compounds. Of these, freon (CCl_2F_2) and sulphur hexafluoride (SF_6) were found to be particularly promising as insulating media. The breakdown characteristics of freon, a gas developed primarily as a refrigerant and which has a vapour pressure at 70° F. of about 85 lb./sq. in., have been studied by Trump, Safford and Cloud (1941), both in the pure state and when mixed with air. It has been found to have a dielectric strength about equal to that of air at three times higher pressure. Its use with electrostatic accelerators is complicated by its gradual decomposition under conditions of ionization and sparking. Sulphur hexafluoride has been

employed by Buechner, Van de Graaff, Sperduto, Burrill, McIntosh and Urquhart (1946) as the insulating medium for the electrostatic accelerator shown in figure 8. It is markedly superior to freon, both because of its much greater chemical stability and because its vapour pressure is about 350 lb./sq. in. at ordinary temperatures (Schumb, 1947).

An increase in the dielectric strength of the gaseous medium favourably affects all of the physical dimensions of the apparatus, including the terminal diameter, the terminal-to-tank spacing, the length of column, and the required belt width and speed for a given generator current. Alternatively, an increase in the dielectric strength of the gaseous medium increases the voltage and current rating of a given generator, as indicated by figure 7, and by the fact that the 2-million-volt generator illustrated in figure 8, but with an added 3 feet of column, attained a voltage of 5·6 million when operated in SF_6 at a pressure of 200 lb./sq. in. In the limit, the volume of electrostatic apparatus of given voltage rating varies inversely as the cube of the dielectric strength. There is little doubt that, except for the expense of the gas and the necessity of storing it in those experimental installations where frequent access to the generator is required, sulphur hexafluoride is to be preferred for the insulation of electrostatic accelerators. This is particularly true when the maximum voltage or greatest compactness is desired.

The increase of breakdown strength with pressure found with uniform electric fields is not realized to the same extent with non-uniform fields. In the extreme case of a positive point directed at a plane, it has been shown by Goldman and Wul (1936 a) and by Howell (1939) that the breakdown strength actually diminishes sharply at a critical pressure which is dependent upon the nature of the gas and its temperature. Above the critical pressure, flashover takes place with an almost negligible amount of predischarged corona current. The critical pressure occurs at about 11 atmospheres for air and in general lies below 30 atmospheres, which is the critical pressure for helium. This positive point-to-plane phenomenon, which can be explained by the space charge surrounding the positive point, is not observed when the polarities are reversed. The positive point-to-plane effect complicates the positive spray and collector problem in high-pressure-insulated electrostatic generators. It may also account for the lower positive terminal voltage occasionally encountered in pressure-insulated generators when the terminal-to-tank geometry produces a highly divergent field.

(ii) *Solid insulators in compressed gases*

High-voltage apparatus insulated by compressed gas requires the use of solid materials for physical supports. In the case of electrostatic accelerators, the possibility of flashover of such solid insulators occurs between the equipotential planes of the column, along the electrically charged belt, and along the outer surface of the acceleration tube. Even in a uniform field the flashover voltage of such solid insulation is generally considerably lower than for the gas-filled gap of equal length.

By proper choice of material and surface contouring the flashover performance of the solid insulator can be brought close to that of the gas in which it is immersed or to its own volume breakdown strength (Goldman and Wul, 1936 b). Figure 9 shows the d.c. flashover strength of contoured cylinders of lucite, textolite, and

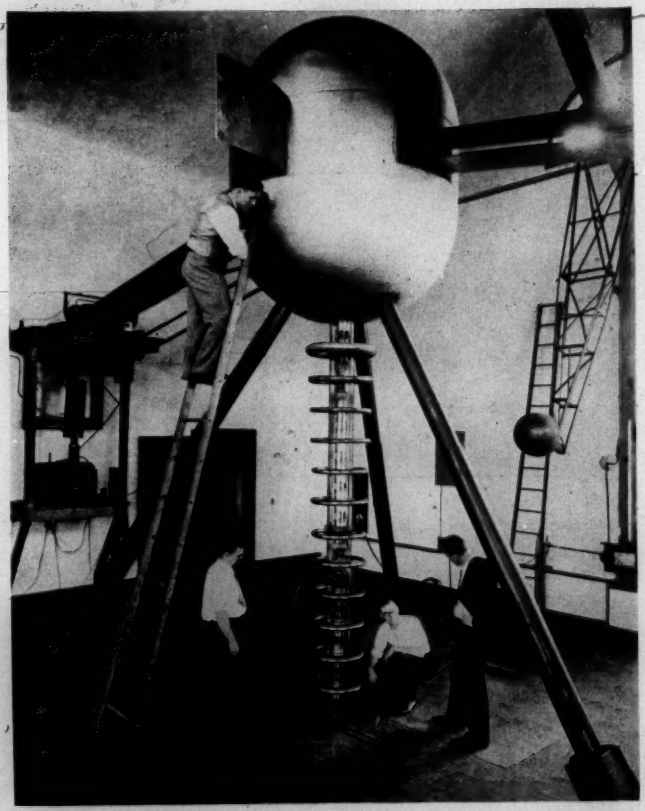


Figure 2. Electrostatic accelerator at the Carnegie Institution of Washington. Dr. M. A. Tuve is at the right.

To face page 8

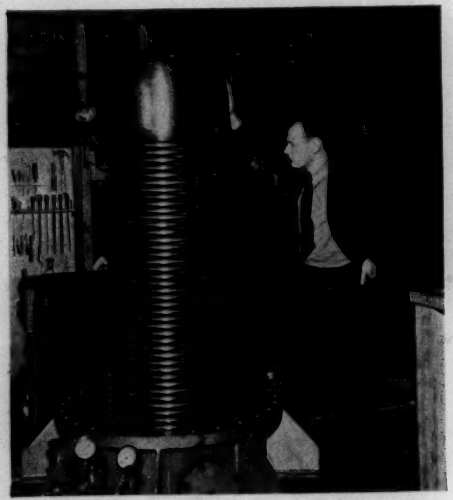


Figure 3. 1·25 Mev. x-ray generator in use for deep therapy at the Massachusetts General Hospital since 1940. Insulated by air at 150 lb./sq. in., the column is 50 in. high, the terminal and inner tank diameters are 24 and 34 in. respectively. The single 14-in.-wide belt and the acceleration tube are within the column, the anode end of the tube extending into the treatment room below.

Figure 5. Pressure-insulated electrostatic accelerator at the University of Wisconsin. Dr. R. G. Herb is in the foreground.

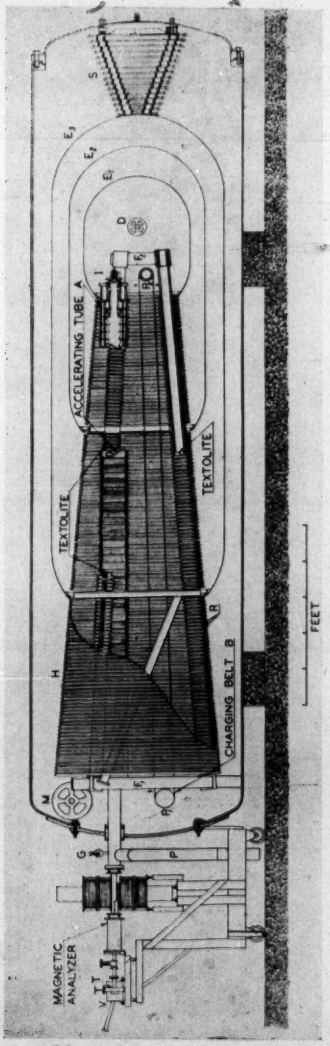


Figure 4. Sectional drawing of the University of Wisconsin electrostatic accelerator, showing horizontal construction and concentric equipotential shields.
(*By courtesy of Dr. R. G. Herb.*)

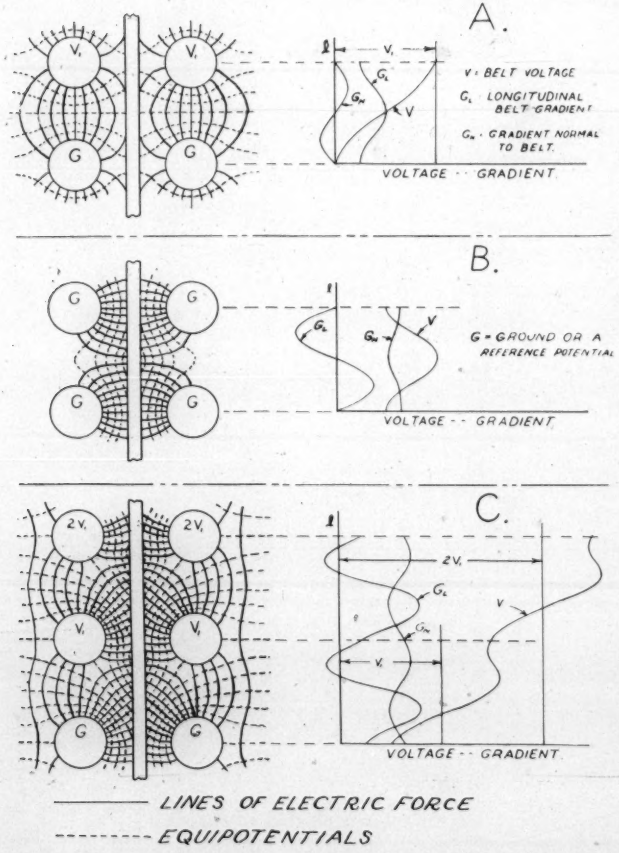


Figure 6. Electric-field picture in vicinity of belt and control rods for the case when (A) belt is uncharged and voltage is applied between adjacent control rods, (B) belt is charged and rods are grounded, (C) normal operating condition with belt charged and voltage applied between rods.

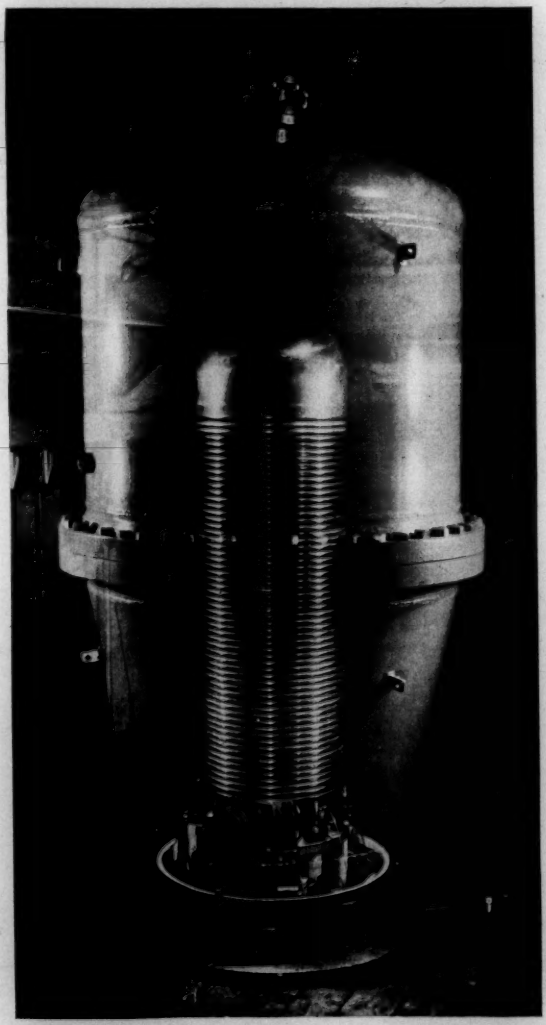


Figure 8. Double-exposure photograph showing insulating column and 200 lb./sq. in. pressure tank of a 2-million-volt accelerator built for industrial radiography.

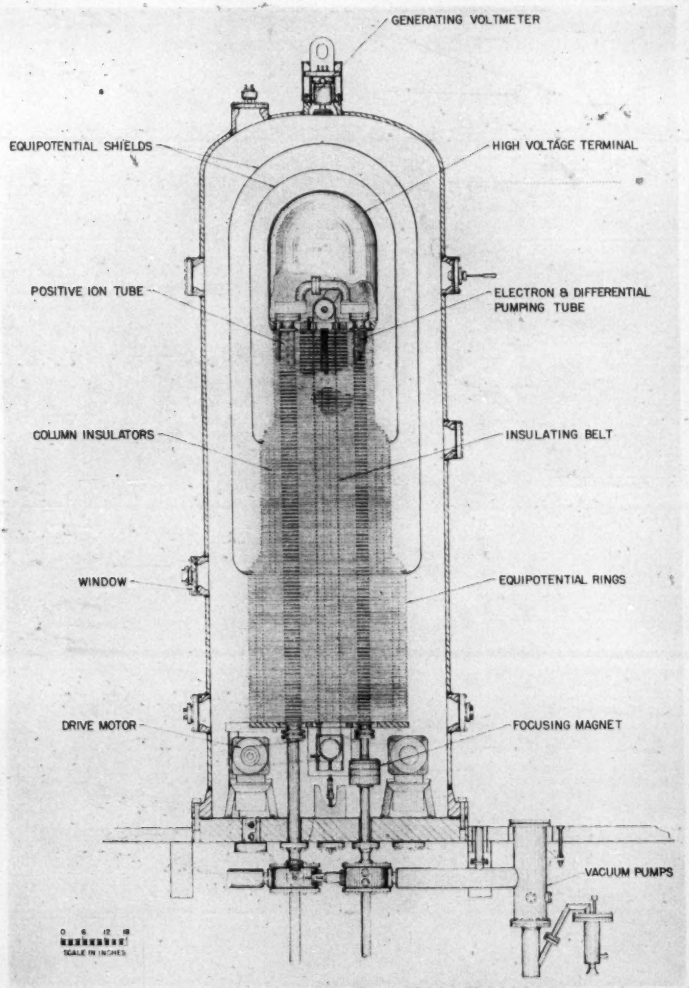


Figure 11. Sectional drawing of accelerator shown in figure 12, showing the two accelerating tubes, one for positive ions and the other for electrons or for differential pumping on the ion source.

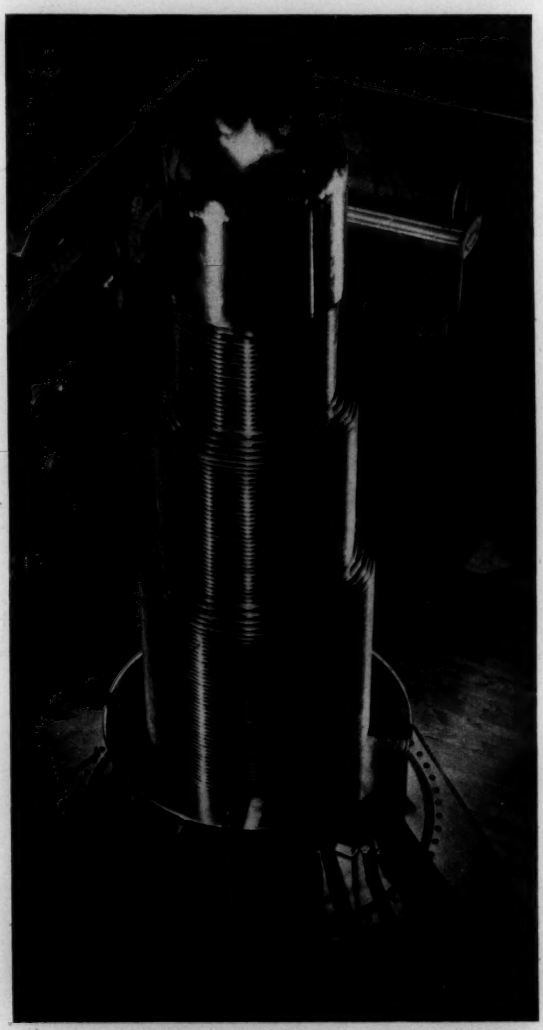


Figure 12. Electrostatic accelerator at the Massachusetts Institute of Technology. Nitrogen and nitrogen-SF_6 mixtures at 400 lb./sq. in. pressure are used as the insulating medium.

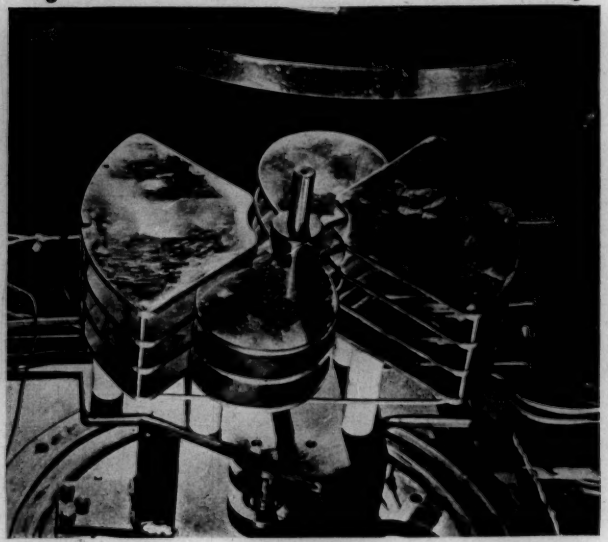


Figure 13. A two-pole synchronous electrostatic generator (or motor) insulated in high vacuum. This machine delivered 60 watts at 75 peak kilovolts of 60 cycle power with an efficiency of over 99 per cent.

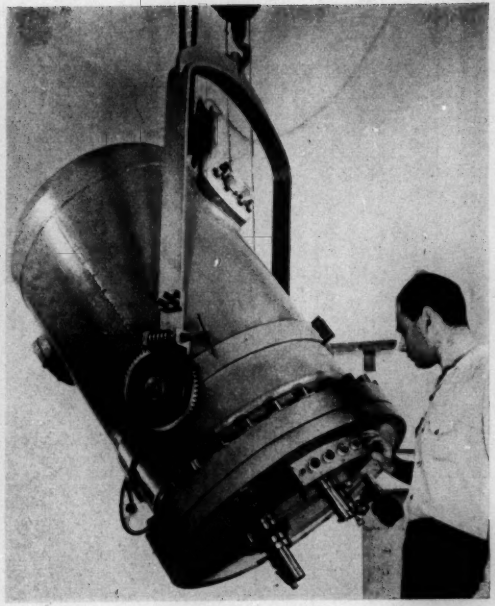


Figure 14. A 2-million-volt electrostatic generator for x rays insulated by nitrogen at 400 lb./sq. in. The insulating column is 2 feet high, the internal dimensions at the tank are 30 × 54 in.

isolantite, immersed in nitrogen at high pressures for the case of a uniform field with a $\frac{1}{2}$-in. gap, and compares these with the flashover strength of the nitrogen gap itself (Trump and Andrias, 1941). It is seen that the flashover strength of corrugated lucite is nearly the same as that of the gas gap, the departure increasing more rapidly the higher the pressure. At a pressure of 100 lb./sq. in. the relative strengths of these materials on the basis of 100 per cent for the nitrogen gap were 80 per cent, 48 per cent, and 40 per cent respectively. The favourable influence of corrugations is the greater the more imperfect the solid dielectric as to surface irregularities, poor electrode-contact conditions, and surface contaminations.

(iii) *Possibilities of vacuum insulation*

Vacuum insulation, which is essential for the acceleration of the charged particles within the accelerating tube, ultimately may also be of value as the insulating medium for the high-voltage generator itself. Should this prove feasible, many of the present problems of accelerator design, such as the increasing weight of the high-pressure enclosures, would be eliminated.

In the low-voltage range below 20 kilovolts the fairly extensive studies of vacuum insulation have shown that breakdown begins with high field emission from the cathode surfaces. This emission begins at fairly definite gradients which depend on the electrode material and its surface condition. At low voltages, gradients as high as 5 million volts per centimetre have been insulated at the cathode surface, and gradients at least an order of magnitude higher than this have been insulated at the anode. The cathode gradient at breakdown in a uniform field diminishes markedly at higher voltages, however. At 100 kilovolts constant potential, the maximum cathode gradient is reduced to about 1 million volts per centimetre and at 700 kilovolts this gradient is reduced to less than 100 kilovolts per centimetre. It has now been established that breakdown in high vacuum at high voltages is not primarily dependent on high field emission as was generally supposed, but involves processes which depend on the total voltage across the inter-electrode gap (Trump and Van de Graaff, 1947). The total-voltage processes which contribute to the interchange of charged particles between cathode and anode include electron bombardment of the anode surface with subsequent emission of positive ions and photons, and ion bombardment of the cathode surface as well as photoelectric absorption with subsequent emission of electrons. All of these processes are evidently dependent upon the total voltage between the electrodes.

Thus, in a vacuum-insulated electrostatic generator it would be necessary, in the present state of technique, to surround the high-voltage terminal with a series of concentric conducting shells which would have the effect of dividing the total voltage into a number of small increments. It would also be desirable to make use of non-organic solid insulation such as glasses and porcelain. This would eliminate the common type of organic charge-conveyor belt and make more feasible a variety of charge-transferring systems composed essentially of spaced metallic carriers. For high-voltage, low-current electrostatic generators, vacuum insulation has not, as yet, been advantageously employed in place of gases at high pressure.

§5. THE ACCELERATING TUBE

(i) *Introduction*

The purpose of the accelerating tube is to provide a region in which charged particles of the desired type can be produced and accelerated to emerge from the tube in a beam, preferably small in diameter and parallel, with an energy corresponding to the generator voltage. To accomplish this, the tube must be highly evacuated so that the charges being accelerated will not suffer collisions in their passage down the tube. Such collisions would not only cause an energy spread in the beam, but might also result in the production of ions in the tube and the scattering of charges out of the beam. Such scattered particles not only represent a loss in beam current, but, more seriously, they may also liberate further charges and neutral gas if they impinge on the tube electrodes or walls. The electric field along the tube must be so arranged both to focus and accelerate the charged particles, keeping the voltage gradient at, and the total voltage between, adjacent metal electrodes within safe values. The influence of the total voltage on breakdown has been mentioned in the preceding section. The field along the insulating walls of the tube must be controlled and these walls must be shielded from any stray ions from the tube which may facilitate flashover or puncture. Higher generator-column gradients resulting from refinements in design and the use of improved insulating media increase the severity of the insulation requirements on the accelerating tube. Gradients of one million volts per foot are now being generated at operating pressures of 400 lb./sq. in. To produce a vacuum tube capable of insulating such voltage gradients and of producing a beam of particles homogeneous in energy requires the most careful design and construction.

(ii) *Multiple-acceleration tubes*

For total voltages higher than a few hundred kilovolts, the adequate control of the electric field along the insulating walls and on the tube electrodes requires that the tube be constructed with alternate conducting and insulating sections, the total generator potential being divided between them. Where average tube gradients of less than a few hundred kilovolts per foot are encountered, as in generators operating either at atmospheric or relatively low pressures, these insulating sections are commonly either porcelain or glass cylinders several inches long and of large diameter. The conducting sections are metal plates placed between these insulating cylinders, the joints being made vacuum-tight by a gasket arrangement. These metal plates are open at their centres to provide a region for the ions to pass down the tube, and in this region the plates carry a cylinder or some similarly shaped electrode. Such electrodes are visible in the tube shown in figure 2. These electrodes shape the electric field along the tube axis so as to accelerate and focus the charged particles. At the same time they are designed to shield the tube wall from stray ions produced by or scattered from the primary beam. The generator potential is distributed along these electrodes either by corona gaps or resistors placed along the outside of the tube. In the case of generators operating at pressures higher than atmospheric, this is commonly done by connecting the tube electrodes to corresponding equipotential planes in the generator column. In this arrangement, the charged particles are not accelerated uniformly down the tube, but gain their energy mainly in discrete steps as they

pass across the gaps between the tube electrodes. The focusing of the charges takes place almost entirely in the first few gaps at the source end of the tube, where the charges are moving slowly. The function of the remaining electrodes in the tube is primarily to distribute the potential and to shield the insulating walls. This shielding reduces the possibility of flashover or puncture of the walls, and shields the beam from the effects of any surface charges which may exist on the insulators.

(iii) *The uniform field accelerating tube*

As the operating gradients along the tube are increased, the flashover of the insulating surfaces on the high-vacuum side becomes the limiting factor in design. This limitation is not generally experienced with tubes which have a normal operating gradient less than 400 kilovolts per foot. It is found that the flashover voltage of a given material in vacuum is substantially lower than when it is immersed in a gas at elevated pressures, and at the higher gradients it is thus desirable to decrease the spacing between adjacent tube electrodes in order to keep the total voltage across the individual insulating sections at a safe value. Fortunately, as the spacing between electrodes is decreased, the insulating walls, for a given tube diameter, are better shielded from the beam. If this separation is decreased sufficiently, the cylindrical element may finally be omitted, and it becomes possible to construct a tube the electrodes of which are simply flat metal discs spaced apart by short sections of insulation. If the total potential difference along the tube is uniformly distributed between these electrodes, a uniform electric field is established along the tube axis and at the electrode surfaces. Thus, the potential gradient at the electrodes is only the average gradient applied to the tube as a whole. Furthermore, the insulating sections are located between closely spaced conducting planes, the potential difference between which is maintained at a moderate value. The appropriate potential distribution along the tube is ensured by connecting appropriate tube electrodes to the potential planes of the generator column. As was discussed previously, these planes are interconnected with resistors so that the generator potential is uniformly distributed among them. These connections between the tube electrodes and the equipotential planes may be made with strips of semi-conducting material, such as neoprene. Such strips have desirable mechanical properties, and if their resistance is of the order of a few megohms, they effectively damp out any high-voltage surges which might otherwise be transmitted from the generator column into the accelerating tube. If, as is often the case in practice, the spacing between adjacent tube electrodes is less than the spacing between the equipotential planes in the generator column, the connection need only be made between the electrode and the plane which is at the same level. The potentials of the intermediate tube electrodes will assume their proper values through leakage and corona.

Apart from their improved insulating properties, accelerating tubes constructed of closely spaced electrodes and having a uniform electric field in the region of acceleration possess many important advantages over the more usual type of tubes which employ cylinders or similarly shaped electrodes for the purpose of so warping the electric field as to produce a strong focusing action on the particle being accelerated. In the uniform-field type of tube, the lines of electric force are

straight lines between the source and the target ends of the tube and there are no transverse components of the field to cause the charged particles to stray off the tube axis. Consequently, if they are introduced into the tube from an electron or ion source which is on the tube axis, their only transverse components of velocity will be those introduced from the source itself. These undesirable components may be minimized by a careful design of the source. It would be expected that the focusing properties of the uniform field would be much less sensitive to irregularities in the potential distribution along the tube than is the case for those geometries in which transverse components of electric field are deliberately introduced. That this is the case in practice is indicated by experience in accelerating electron beams. For example, with a tube of the type shown in figure 10, and no especial effort made to ensure an even potential distribution down the tube, it is found that electrons emitted from a small plane cathode placed on the tube axis are accelerated into a beam, the shape of which corresponds to that of the emitting area and the cross-sectional dimensions of which may be approximately calculated from a

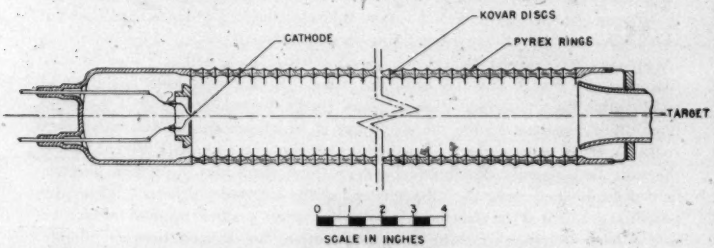


Figure 10. Cross-section drawing of uniform field accelerating tube.

knowledge of the cathode dimensions and the thermal spread of velocity of the electrons. When the beam emerges from the region of acceleration, it has approximately the cross-section corresponding to the cathode area and, in the case of a 57-in.-long tube operating at two million volts, has a maximum dimension of approximately $\frac{1}{4}$ in. This is approximately the spread which would be calculated from the initial velocity spread of the electrons from the cathode. It is found that the position of the beam and the current is independent of the total tube voltage over the range from 300 kilovolts to 3 million volts. This independence of position and current on voltage is a particular advantage when the accelerator is used for problems involving measurements as a function of bombarding energy or where the accelerator is pulsed to obtain high currents of short duration. When tubes employing strongly convergent fields are used over this voltage range, much more attention must be given to the voltage distribution between the electrodes, particularly at the source end of the tube.

As the beam of charged particles emerges from the region of acceleration, it is, because of the steady applied potential, quite homogeneous in energy and, because of the acceleration in the uniform field, quite parallel in direction. The energy distribution in the beam is only the spread introduced by the source of the charged particles. In the case of positive ions, this makes for great ease and precision of any

subsequent magnetic or electrostatic analysis of the beam. In the case of electrons, such beams may be further focused by magnetic lenses following the technique developed for electron microscopes. In this way, electron beams of several hundred microamperes have been produced which have diameters less than 0·010 in. (Buechner, Van de Graaff, Sperduto, McIntosh and Burrill, 1947).

It is found that even with tube pressures somewhat higher than 10^{-5} mm. Hg, electron beams of the order of a milliampere may be accelerated into a well-defined beam. Consideration of the relative probabilities of ionization by high-energy electrons and positive ions leads to the conclusion that, for a given amount of ionization in the accelerating tube, the tube pressure must be lower than this by a factor of approximately 100 if positive-ion beams of this intensity are to be obtained. Thus, when beams of positive ions are to be produced, provision must be made for pumping away the neutral gas from the ion source. Recent developments in ion sources have indicated the advantages of removing most of this neutral gas by differential pumping at the ion source (Lamar and Buechner, 1947). The rugged construction and excellent insulating properties of the uniform field-type tube having a large number of alternate insulating and conducting sections indicate that a separate tube could be used for such pumping, the pump for this purpose being located at ground potential and out of the high-pressure region. The fact that the generator polarity may be readily changed makes it feasible to provide this second tube with an electron source. Thus, several electrostatic accelerators are now being constructed (Van de Graaff, Buechner, Woodward, McIntosh, Burrill and Sperduto, 1946) which will have two vacuum tubes, one to be used for accelerating positive ions and the other for accelerating electrons or for differential pumping on the ion source, depending on the generator polarity. The general design of these accelerators was made possible by a grant from the Rockefeller Foundation to the Massachusetts Institute of Technology. The one under construction at that institution is shown in figures 11 and 12, the first being a sectional drawing which shows the arrangement of the two vacuum tubes.

(iv) *Accelerating-tube construction*

A principal problem in the construction of accelerating tubes to be used in equipment operating at high gas pressures lies in the necessity of providing very reliable vacuum-tight joints between the metal electrodes and the insulating sections. This is particularly true with the more recent accelerator designs where tubes with three electrodes per inch and of over 100 inches in total insulating length are employed and where the insulating gas is at a pressure of 400 lb./sq. in. Exceedingly small leaks can be very troublesome, particularly when positive ions are accelerated, since very slight traces of air or the other heavier gases used as the insulating medium surrounding the tube are as effective in scattering the beam of charged particles as much higher pressures of hydrogen. Under these conditions, the use of gaskets at each joint is not practical, both because of their uncertain reliability and the mechanical problem of assembly.

There are at present two principal methods for the construction of tubes which are to operate under these high pressures and voltage gradients. In one, the joints are made vacuum-tight through a glass-blowing operation. Here the metal electrodes must be either of a material such as Kovar, the coefficient of thermal

expansion of which matches that of the glass used as the insulating sections, or they must be thin and of a material suitable for sealing to the insulators according to the Housekeeper (1923) technique. Because of space limitations in accelerators operating at high pressures and the desirability of minimizing the stresses on the tube due to these pressures, tubes are made with as small a diameter as is consistent with adequate vacuum pumping speeds. A cross-section drawing of a Kovar-glass tube with an outside diameter of three inches and a spacing between electrodes of one third of an inch is shown in figure 10. This tube is produced commercially (Machlett Laboratories, Inc.) and is used, sealed off from the vacuum pump, with generators constructed as shown in figure 8, for the production of 2-million-volt X rays. Tubes constructed in this way are very clean from the vacuum point of view since they contain no sources of organic vapours which might be particularly troublesome when the tubes are used for the acceleration of positive ions. They suffer from the disadvantage that the choice of insulating material is restricted, and this limitation may be serious when the highest tube gradients are required.

A study of the flashover of solid insulators in high vacuum has shown both the marked dependence of flashover voltage on the nature of the insulator and the value of surface contouring. It has been found that the various electrical porcelains are low in volume puncture strength and are generally inferior to the hard glasses as regards surface flashover. The highest overall performance has been found with a 96 per cent silica glass known as Vycor (Corning Glass Works No. 790). This material, in a typical contoured test section one centimetre long, regularly insulated 100 kilovolts after conditioning, and six such sections in series have insulated 500 kilovolts across a total column length of 2¼ in., including the tube diaphragms. This is higher by a factor of two over the performance of Pyrex glass (Corning), a material often used in tube construction. Much more flexibility in the choice of both electrode and insulating materials is afforded by a technique for cementing them together with a thermo-plastic such as plasticized polyvinyl acetate, commercially available under the trade name Vinylseal (Carbide and Carbon Chemicals Corp.). Joints made in this manner are exceedingly vacuum-tight and are remarkably strong mechanically. A principal problem with this method lies in the desirability of keeping the pressure due to organic vapours low in the high vacuum region. This may be accomplished by careful joint construction and assembly.

§ 6. VOLTAGE MEASUREMENT

If conditions in the accelerating tube are such that the charged particles being accelerated do not lose energy in collisions, then, at any instant of time, the beam from an electrostatic accelerator is homogeneous in energy, the only energy spread being that introduced by the ion or electron source. Thus this energy may be determined either by measurements on the beam or by measurements of the generator voltage. The latter is simpler in practice, and in many cases such measurements can be made with sufficient precision that it is not necessary to analyse the beam to determine its energy.

While in principle the most direct method for determining the generator voltage consists in measuring the current through a known resistor, in practice it is not possible to obtain resistors of sufficient accuracy and of high enough value and to shield them from corona effects, particularly when the voltage to be measured

is over 1 million volts. Consequently, the voltage of the generator is usually measured by determining the electric field it produces, using a generating voltmeter. Such a voltmeter is placed as indicated in figure 11 and consists of a number of insulated conductors which are alternately exposed to and shielded from the electric field by means of a rotating grounded conductor. The resultant alternating currents which are introduced in the circuit to which these insulated conductors are connected are rectified either mechanically (Parkinson, Herb, Bernet and McKibben, 1938) or electrically (Trump, Safford and Van de Graaff, 1940), and the direct current thus obtained is a measure of the field and hence of the generator voltage. As would be expected, it is found that the voltmeter output is linear with generator voltage, and it may be calibrated by applying to the high-voltage terminal a voltage which is low enough to be measured accurately with wire-wound resistors. A more reliable calibration may be obtained by operating the accelerator in its working range and utilizing effects produced by the beam to determine the voltage at certain points. Thus, when electrons are being accelerated, convenient calibration phints are provided by the thresholds for photodisintegration of elements such as deuterium and beryllium (Myers and Van Atta, 1942). In the case of positive ions, the thresholds for various nuclear reactions and resonances are used (Hanson and Benedict, 1944). Once calibrated, such generating voltmeters are simple and convenient to use and may be relied on to determine the voltage to somewhat better than one per cent.

Where higher accuracy is required, or where the intermediate electrodes previously referred to are employed to increase the maximum voltage of the generator, the voltage is determined by deflection of the beam. This is done using either electrostatic or magnetic fields, and the field required to cause a certain deflection is a measure of the energy. In either case, the analyser may be calibrated by measuring the field required to deflect charged particles of known energies, as those from radioactive sources, or in terms of the excitation by the deflected beam of known nuclear resonances or reactions. The electrostatic method has the advantage that, once calibrated, the voltage between the plates is a direct measure of the energy (Hanson, 1944). In working with positive ion beams, it is in general necessary to analyse the beam in mass. Since the deflection measurements can be made on the undesired beam component, the analysis as to energy does not necessarily require a separate deflecting field.

§7. VOLTAGE STABILIZATION

The voltage as generated is sufficiently steady for many purposes, and manual control can be used to eliminate slow drifts. These slow drifts may also be eliminated if a circuit is employed to vary the charging current sprayed on the belt so as to keep the terminal voltage constant. This may be done either by controlling the primary voltage of the high-voltage transformer (Hanson, 1944) or by varying the grid potential of a small pentode placed in the grounded lead of the rectifier set. In the latter case, the potential drop across the pentode is in series with that between the spray points and the lower grounded pulley and hence, for any given rectifier voltage, variation of this drop varies the spray current. This has the advantage that the control circuit is not required to deliver power. In either case, the controlling signal is derived from a voltage-measuring device. Such arrangements of

course are not capable of correcting for voltage variations which have a frequency higher than that corresponding to the speed of the charging belt. However, because of the smoothing effect of the terminal capacitance, this is not so serious a limitation as it might at first appear.

Faster control can be obtained by placing an array of sharp points so as to spray corona to the high-voltage terminal. If the electric field at the points is modified by an adjacent electrode (Ashby and Hanson, 1942), whose potential can be varied by appropriate connection to the means of voltage measurement, the speed of voltage control is limited only by the time of flight of the ions between the points and the terminal and by the terminal capacitance. Since the generator voltage is linear with the load current, and since the corona current can be made to vary rapidly with the potential of the control electrode, it is possible to obtain very stable operation in this way, particularly when the voltage measurement is done by deflection of the beam (McKibben, Frisch and Hush, 1946).

When the generator is provided with the concentric shields previously mentioned, the above method is still applicable since a variation in the potential of the outermost shield is accompanied by the same variation in the potential of the terminal which it encloses. Here the large capacitance of the shield tends to slow down the speed of control because of its large charging time. The situation may be improved by providing an insulated shield which encloses essentially the entire column and terminal assembly and which is normally at ground potential. The potential of this shield can then be controlled by conventional circuits connected to the means of voltage measurement.

It is also possible when operating with positive polarity to stabilize the accelerating voltage by arranging to send an electron beam through the accelerating tube or through the auxiliary vacuum tube which may be provided for differential pumping. The magnitude of the electron current may be controlled by a grid appropriately connected to the energy-measuring device. Where this is done, it may be desirable to isolate the ion source from the remainder of the terminal in order to reduce the capacitance which must be charged by the electron current. With such a stabilizing circuit, Bennet, Bonner, Mandeville and Watt (1946) have reduced to ± 150 volts the potential fluctuations of an electrostatic accelerator operating at 1 million volts.

§8. OTHER TYPES OF ELECTROSTATIC GENERATORS

Although the flat belt travelling in a gaseous medium has thus far been the most practical charge-conveyor for high-voltage electrostatic generators, other carriers have been considered and investigated by many laboratories. Among these other methods are the use of an insulating fluid such as oil circulated through an endless pipeline by a pump, a whirling insulating disc either carrying charge directly or carrying charged metallic carriers, and spaced charged particles such as dust, droplets, or metallic spheroids. The current capacity of these various carriers depends, as with the belt, on the dielectric strength of the medium itself, on the surface area enclosing the charged carriers, and on the linear speed of the carriers with respect to the terminal.

It is possible to devise many forms of a.c. and d.c. electrostatic generators which are essentially variable condensers with interleaving metallic plates. Such

generators operated in high vacuum, while suitable primarily for intermediate voltages, have been investigated by Trump (1947) because they offer a possibility for the generation of electrical power at high voltage. A constant potential characteristic may be obtained either by rectifier circuits or by commutation at appropriate times in the charging cycle. Figure 13 shows a vacuum-insulated variable-condenser generator which has operated as either a synchronous motor or generator with high output efficiency. If gradients of several million volts per centimetre can be insulated in high vacuum at suitably high voltages, then such generators may attain compactness comparable with that of electromagnetic power machinery. This combination of high voltage and high gradient is not yet feasible, but may be realized as the result of further study of vacuum insulation.

§9. PRESENT TRENDS IN THE DESIGN OF ELECTROSTATIC ACCELERATORS

The increasing interest in pressure-insulated electrostatic accelerators for nuclear research and for medical and industrial applications is directing present development efforts towards higher voltages and greater compactness as well as the industrial production of such accelerators. Many laboratories are now engaged in the construction of large electrostatic accelerators. At the Massachusetts Institute of Technology the general results on electrostatic design and the insulation of high voltages in compressed gases and in high vacuum are being embodied in the design of an accelerator for both positive ions and electrons which will operate in the voltage range up to about 12 million volts. The final design of this accelerator, as well as the determination of its actual upper voltage limit, will be influenced by experience with the equipment shown in figure 12 and by the results of research now under way on the insulating capabilities of solid insulation in compressed gases and high vacuum, and on such problems as the effect of discharges on the belt and tube surfaces.

Probably the most compact accelerator developed so far is the one illustrated in figure 14, which has been developed at the Massachusetts Institute of Technology as a 2-million-volt x-ray and cathode-ray source for deep therapy, for industrial radiography, and for scientific purposes. By refinement of electrostatic design and by the use of permanent gases at 400 lb./sq. in. pressure, this generator can operate with a normal column gradient of 1 million volts per foot and with a terminal-to-tank spacing of 7 in. The normal electron-beam current at full voltage is 250 microamperes. In this design both the charged conveyor belt, which is 6 in. wide and which travels at 4000 feet per minute, and the sealed-off acceleration tube are the limiting factors and are enabled to operate at these elevated gradients by careful control of the electric fields and choice of insulating material.

REFERENCES

ASHBY, R. M. and HANSON, A. O., 1942. *Rev. Sci. Instrum.*, **13**, 128.
BARTON, H. A., MUELLER, D. W. and VAN ATTA, L. C., 1932. *Phys. Rev.*, **42**, 901.
BENNET, W. E., BONNER, T. W., MANDEVILLE, C. E. and WATT, B. E., 1946. *Phys. Rev.*, **70**, 882.
BUECHNER, W. W., VAN DE GRAAFF, R. J., SPERDUTO, A., BURRILL, E. A., MCINTOSH, L. R. and URQUHART, R. C., 1946. *Phys. Rev.*, **69**, 692.

BUECHNER, W. W., VAN DE GRAAFF, R. J. SPERDUTO, A., MCINTOSH, L. R. and BURRILL, E. A., 1948. *Rev. Sci. Instrum.*, **19** (in press).
CHARLTON, E. E. and COOPER, F. S., 1937. *Gen. Elect. Rev.*, **40**, 438.
GOLDMAN, I. M. and WUL, B. M., 1936 a. *J. Tech. Phys., U.S.S.R.*, **3**, 16.
GOLDMAN, I. M. and WUL, B. M., 1936 b. *J. Tech. Phys., U.S.S.R.*, **3**, 519.
HANSON, A. O., 1944. *Rev. Sci. Instrum.*, **15**, 57.
HANSON, A. O. and BENEDICT, D. L., 1944. *Phys. Rev.*, **65**, 33.
HERB, R. G., PARKINSON, D. B. and KERST, D. W., 1937. *Phys. Rev.*, **51**, 75.
HERB, R. G., TURNER, C. M., HUDSON, C. M. and WARREN, R. E., 1940. *Phys. Rev.*, **58**, 579.
HOUSEKEEPER, W. G., 1923. *Elect. Engng.*, **42**, 954.
HOWELL, A. H., 1939. *A.I.E.E. Trans.*, **58**, 193.
JOLIOT, F., FELDENKRAIS, M. and LAZARD, A., 1936. *C.R. Acad. Sci., Paris*, **202**, 291.
LAMAR, E. S. and BUECHNER, W. W., 1947. *J. Appl. Phys.*, **18**, 22.
MCKIBBEN, J. L., FRISCH, D. H. and HUSH, J. M., 1946. *Phys. Rev.*, **70**, 117.
MYERS, F. E. and VAN ATTA, L. C., 1942. *Phys. Rev.*, **61**, 19.
NATTERER, K., 1889. *Ann. Phys. Chem.*, **38**, 663.
PARKINSON, D. B., HERB, R. G., BERNET, E. J. and MCKIBBEN, J. L., 1938. *Phys. Rev.*, **53**, 642.
POULSEN, G. G., 1940. Thesis (Mass. Inst. Tech.).
RODINE, M. T. and HERB, R. G., 1937. *Phys. Rev.*, **51**, 508.
RUTHERFORD, E., 1927. *Proc. Roy. Soc., A*, **117**, 310.
SCHUMB, W. C., 1947. *Ind. Engng. Chem.*, **39**, 441.
TRUMP, J. G., 1947. *Elect. Engng.*, **66**, 525.
TRUMP, J. G. and ANDRIAS, J., 1941. *A.I.E.E. Trans.*, **60**, 986.
TRUMP, J. G. and CLOUD, R. W., 1943. *Am. J. Roentgenology and Radium Therapy*, **49**, 531.
TRUMP, J. G., SAFFORD, F. J. and CLOUD, R. W., 1941. *A.I.E.E. Trans.*, **60**, 132.
TRUMP, J. G., SAFFORD, F. J. and VAN DE GRAAFF, R. J. 1940. *Rev. Sci. Instrum.*, **11**, 54.
TRUMP, J. G. and VAN DE GRAAFF, R. J., 1939. *Phys. Rev.*, **55**, 1160.
TRUMP, J. G. and VAN DE GRAAFF, R. J., 1947. *J. Appl. Phys.* (in publication).
TUVE, M. A., HAFSTAD, L. R. and DAHL, O., 1935. *Phys. Rev.*, **48**, 315.
VAN DE GRAAFF, R. J., 1931. *Phys. Rev.*, **38**, 1919.
VAN DE GRAAFF, R. J., BUECHNER, W. W., FESHBACH, H., SPERDUTO, A., BURRILL, E. A. and MCINTOSH, L. R., 1946. *Phys. Rev.*, **69**, 46.
VAN DE GRAAFF, R. J., BUECHNER, W. W., WOODWARD, W. M., MCINTOSH, L. R., BURRILL, E. A. and SPERDUTO, A., 1946. *Phys. Rev.*, **70**, 797.
VAN DE GRAAFF, R. J., COMPTON, K. T. and VAN ATTA, L. C., 1933. *Phys. Rev.*, **43**, 149.

RADIOACTIVE BRANCHING

By N. FEATHER, F.R.S.,
Department of Natural Philosophy, The University, Edinburgh

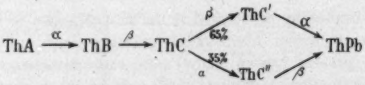
§1. INTRODUCTION

If radioactive branching be defined as the susceptibility to mutually exclusive modes of disintegration, whereby recognizably distinct products result, then there are at present known many more radioelements which exhibit this phenomenon than the three C-bodies (of the radium, thorium and actinium series) with which it was first discovered and studied in some detail. It will be the object of this report to classify the more recently discovered cases and to discuss the whole phenomenon in general terms.

The suggestion that more than one type of disintegration might be possible for a given type of radioactive atom appears to have originated with Soddy (1909). Two quotations from this paper will show the clearness with which the fundamental issues were grasped and the possibilities foreseen, within a few years of the formulation of the disintegration hypothesis in its original form. "The question about to be discussed", Soddy wrote, "is whether necessarily only one mode of instability can exist within the atom at the same time". And, having rejected this condition as a necessary limitation, "It follows that the question by which particular mode any individual atom will disintegrate remains indefinite up to the moment of disintegration and no process is conceivable which would sort a collection of atoms according to the mode by which eventually they will disintegrate and to the product which eventually they will produce". The phraseology is that of the pre-nuclear epoch in physics, but the basic ideas are essentially modern. At the time the suggestion was put forward no obvious example of dual or multiple disintegration could be cited, but within five years it had been shown that the transformations of the bodies radium C, thorium C and actinium C were all of this type.

It is an essential and well-known feature of sequences of transformations that, under equilibrium conditions, each active body in a radioactive series in which there is no branching contributes the same number—or very nearly the same number—of disintegrations per unit time. Thus in an unchanged uranium mineral the rates of emission of α-particles from the uranium I, the uranium II, the ionium and the radium contained in the mineral are all the same to the accuracy of measurement. It was Fajans (1911) who first showed that it was possible to isolate a product of the radium series which in equilibrium contributed many fewer than the "normal" number of disintegrations. By recoil from radium C he obtained a β-active body of 1·38 min. half-value period which he postulated as a "branch" product, just on account of its low activity ($\sim 3.10^{-4}$ of the normal: (Fajans, 1912)). The details of a similar branching with thorium C were unravelled much more completely by Marsden and Darwin (1912) and Marsden and

Wilson (1913) very shortly afterwards. In this case the most obvious abnormality was a feature of the α-emission. It had been noted by Hahn (1906) as early as 1906 that two groups of α particles, one about twice as intense as the other, appeared to be emitted by thorium C. On the other hand, Bronson (1908) had shown that these two groups together contribute only as many particles as are provided by either of the single groups from thoron or thorium A (when all three bodies are in equilibrium with radiothorium). A completely consistent description of the "abnormalities" was eventually provided by the now generally accepted branching scheme



..... (i)

The product thorium C'' (half-value period 3·2 min.) could be obtained separately by recoil, in complete analogy with the short-lived β-active body of Fajans (radium C''), but in this case the branching ratio was more nearly unity and there was no cause for suspicion that the recoil process involved was other than α-recoil (the shorter-range, less abundant, α-particles of the thorium C complex being effective). The discovery of a second group of α-particles of small abundance from actinium C (Marsden and Perkins, 1914) then led to the natural assumption that schemes similar to (i) must apply to the transformations of radium C and actinium C as well as to that of thorium C, and the work of Albrecht (1919) carried the experimental verification of this assumption as far as was possible with the methods then available. Complete verification was impossible at the time, but later investigations have provided it in full, and the α/β branching ratios deduced by Albrecht have proved to be reasonably correct. These ratios were: Th C 35/65, Ra C $4/10^4$, Ac C $10^4/16$. Recent values for the ratios are ThC 33·7/66·3 (Kovarik and Adams, 1938), RaC 2·07/10^4 (Rutherford, Wynn-Williams, Lewis and Bowden, 1933) and Ac C $10^4/27$ (Mercier, 1926; Rutherford, Wynn-Williams and Lewis, 1931).

During the ten-year period 1919–29 very little further advance was made, but Fajans (1926) pointed out what was implicit in the original discussion of Soddy, that the partial disintegration constant, λ_α or λ_β, corresponding to whichever disintegration mode is in question, is the significant probability constant to employ in respect of the fundamental nuclear process involved. In particular, in relation to the Geiger-Nuttall rule, partial constants, λ_α, should be used for the competing α-modes of disintegration of the C-bodies of the three series. On the basis of this assertion the difficulty experienced in detecting the α-particle emission of radium C was easily understood: the problem resolved itself into the detection of a group of α-particles having a predicted range of about 4 cm., present to the extent of a few hundredths of one per cent only, when compared with the co-present group of 7-cm. range from radium C'.

A new attack upon the problem of α-particle group analysis began in 1929, and in 1931 the whole problem of β-disintegration acquired a definiteness and an urgency not previously appreciated. The attack upon the α-particle problem arose from the discovery of the phenomenon of "fine-structure" in α-particle emission (Rosenblum, 1929). To this brief reference will have to be made at a later stage;

for the present it is sufficient to note that the new methods resulted in a more accurate determination of particle energies in general and also that they led directly to the discovery of the "missing" radium C α-particles with very closely the energy which had been predicted for them (Rutherford, Ward and Wynn-Williams, 1930; Rutherford, Wynn-Williams, Lewis and Bowden, 1933).

The problem of β-disintegration acquired definiteness through the neutrino hypothesis of Pauli (*see* Carlson and Oppenheimer, 1931), this hypothesis involving the assumption that the limiting energy of the continuous β-particle energy spectrum is a measure of the kinetic energy liberated in the disintegration in all cases—that is, liberated in the form of electron energy plus neutrino energy in each disintegration, for a particular species of nucleus, and always assuming that the same energy state of the product nucleus is involved throughout. A study of the limiting energies of the β-particle spectra of thorium C and thorium C'' (Henderson, 1934) became, therefore, an important investigation relative to the branching process for the former body: but for the complication due to the emission of intense γ rays by thorium C'' (of which the energy relationships are not even yet completely understood) it would have provided a more conclusive proof than in fact it does provide for the equality of the total energy set free in the alternative sequences of changes leading from Th C to Th Pb (see (i)). As it is, Henderson's investigation cannot be regarded as more than the first clear circumstantial evidence that the total energies are the same, or, stated otherwise, that the final product Th Pb is obtained in the same energy state whether through the Th C' or through the Th C'' branch of the thorium series.

The explanation of the only other case of radioactive branching discovered before 1930 might have been given about that time in the statement that with another body the product nucleus is not formed in the same energy state in all cases—coupled with the additional statement that the higher energy state resulting from a certain fraction of the disintegrations is of "abnormal" properties. The case then in point was the so-called β/β branching of uranium X_1, discovered by Hahn (1921). The suggestion had been made some years earlier by Soddy (1917) that nuclear species characterized by the same mass and charge numbers might exist which were radioactively distinct. It appeared from Hahn's results that the two β-emitters, uranium X_2 (half-value period 1·14 min.) and uranium Z (half-value period 6·7 h.), each of which seemed to be formed directly from uranium X_1 (by β-disintegration), were two such isomeric species. This conclusion received further support from several later investigations, in particular, from the work of Walling (1931, 1932).

The idea of relatively long-lived excited states of nuclei which is required to explain the isomerism evident in cases such as this, has had a chequered history. About twenty years ago an experiment of Jacobsen (1927) appeared to indicate a lifetime of the order of 10^{-5} sec. for the γ-ray emission which follows the β-disintegration of radium C, and influenced in part by this interpretation the writer shortly afterwards suggested (Feather, 1929) the need to keep in mind the possibility that γ-ray transition probabilities "as small as $1\,\text{sec.}^{-1}$" might sometimes be found. Then Jacobsen's experimental result was shown to be spurious (Joliot, 1930; Wright, 1932) and it was not until other cases of isomerism were discovered amongst the "artificially produced" radioelements (Szilard and

Chalmers, 1935; Kurtchatow, Kurtchatow, Myssowsky and Roussinow, 1935) that a serious theory of the metastable nuclear state was advanced (Weizsäcker 1936).

It will be clear that when nuclear isomers result from alternative modes of a previous radioactive transformation, not being formed directly as the primary products of induced nuclear change, we may properly speak of radioactive branching in relation to the common parent species, though the definition from which we have started, which requires the production in the branching process of products which are "recognizably distinct", will obviously need further interpretation in regard to this qualifying clause in certain cases. All excited states of nuclei must clearly be regarded as having a finite lifetime in respect of radiative transition (Chalfin, 1929), and estimates may be made of such lifetimes, even when they are of the order of 10^{-12} sec., in favourable cases (Rutherford, Lewis and Bowden, 1933). However, we should certainly disallow classification as branching to the mutually exclusive modes of disintegration of a parent species which result in a product in one or other of a number of nuclear states of lifetime as small as this. If we did not we should have to include all cases of α-disintegration characterized by a "fine-structure" spectrum and every case of β-disintegration in which the continuous energy spectrum is complex. We should have to include, in fact, most of the cases of radioactive disintegration in which both particles and γ-ray quanta appear, as associated radiations. But, if we allow classification as branching to a case in which the lifetime of the "metastable" nuclear state is 1 sec. and disallow it when the state has a lifetime of 10^{-12} sec., the problem of exact definition still remains over a vast range of possibilities. Evidence that this problem is not merely an academic one is provided by the recent report of the discovery of a nuclear state (with $^{181}_{73}$Ta) of 22 microsec. half-value period (De Benedetti and McGowan, 1946). This cannot long remain an isolated case*—and, if it be thought too hastily that lifetimes as small as this should be disallowed in our definition, it should be remembered that thorium C′, of lifetime only one hundredth as long, is accepted as a "recognizably distinct product" in the series of radioelements. The problem of exact definition is never easy for the empiricist!

§ 2. α/β BRANCHING: RESULTS OF RECENT INVESTIGATIONS

The first additions to known cases of α/β branching came as a result of attempts to detect the formation of radioactive isotopes of elements 87 and 85 as products of the "classical" radioelements. As early as 1933 Guében reported certain rather vague indications of the emission of α-particles in a very small fraction of the disintegrations of $^{228}_{89}$MsTh$_2$ (Guében, 1933). These observations have not, however, been substantiated, and Rona and Schintlmeister (1938) set an upper limit of 2×10^{-4} to the α/β branching ratio in this case. In a general discussion in the same year Schintlmeister (1938) suggested that branching was more likely to be observed with $^{227}_{89}$Ac than with the heavier isotope, and this, in fact, is what was discovered by Perey (1939), apparently in ignorance of Schintlmeister's suggestion.

* Bittencourt and Goldhaber (1946) have more recently reported a metastable state of $(5 \pm 2) \times 10^{-6}$ sec. period with $^{121}_{52}$Te and De Benedetti and McGowan (1947) a state of about 10^{-6} sec. period with $^{187}_{75}$Re.

The branching ratio was estimated at about 8/1000 (*see also* Hahn and Rothenbach, 1913; Meyer, Hess and Paneth, 1914) and the branch product $^{223}_{87}$Ac K was identified as a β-active body of half-value period 21 minutes.

So far as element 85 is concerned, the formation of two active isotopes in rare (β) modes of disintegration of the predominantly α-active species radium A and thorium A has more recently been advanced as the most natural explanation of the observation of weak long-range α-particle groups with these species (Karlik and Bernert, 1943). If this interpretation is accepted, α/β ratios of $10^4/4$ and $10^4/1{\cdot}35$ must be recorded. In the case of thorium A, however, there is considerable difficulty in accepting the interpretation. As pointed out independently by Flügge and Krebs (1944) and by the writer (Feather, 1945), since in this case the energy of the newly observed α-particle group of feeble intensity is greater than the sum of the energies of β-disintegration of thorium A and β-disintegration of thorium B (due allowance being made for the energy of γ-emission), it is energetically impossible that it should represent the α-emission of ($^{216}_{85}$) formed in the β-disintegration of $^{216}_{84}$Th A. In the writer's view, if the observations themselves are confirmed, the most likely explanation is in terms of a "long range" group of particles emitted by thorium A itself. The energy criterion applied to the suggested branching of radium A does not expose any parallel difficulty, in this case α/β branching can be accepted without serious question—and the energy of β-disintegration of the newly discovered ($^{218}_{85}$) may be taken as 6·75 Mev. On this basis a half-value period of a few seconds may be predicted for the new species.

Two further cases of indubitable α/β branching must now be added to complete our survey; both of these involve active isotopes of bismuth ($Z=83$). In the "$4n+1$" (neptunium) series of radioelements, of which the artificially produced $^{237}_{93}$Np is the long-lived parent, the C-product ($^{213}_{83}$) is found to exhibit the effect, as do the C-products of the three "classical" series (English, 1945; Segrè, 1946). The α/β branching ratio is about 2/100 and the energy of α-disintegration of the C-product about 6·0 Mev. Secondly, radium E ($^{210}_{83}$) has recently been shown to suffer dual disintegration as suspected by the writer (Feather, 1946). The branching ratio in this case is of the order of $3/10^7$ (Broda and Feather, 1947) and the product of the rare β-mode of disintegration is the 4·2-min. thallium isotope $^{206}_{81}$Tl to which the 3·5-y. activity obtained by the action of slow neutrons on thallium has previously been erroneously assigned (Fajans and Voigt, 1941).

Having presented the facts we are now in a position to review the general topic of α/β branching, so far as the heavy elements are concerned. A partial review was given by Turner (1940), but certain of his conclusions are no longer tenable, and much new information has been obtained since his paper was written. Then a considerable advance was made by Berthelot (1942 a and b), who showed that an empirical regularity made it possible to predict with considerable accuracy the energy of α-disintegration of hypothetical species by a method of interpolation. Independently the writer introduced the same empirical method somewhat later (Feather, 1945), and used it specifically for a discussion of branching (Feather, 1947). The empirical method consists in classifying nuclear species according to isotopic number, $A-2Z$ (since the isotopic number of $^4_2$He is zero, successive members of a sequence of α-active species have the same value of $A-2Z$), and plotting the α-disintegration energy E against the charge number Z for $A-2Z$

constant. In this way a family of similar curves is obtained which may be interpolated, and to some degree extrapolated, with reasonable assurance.

We may take the thorium series bodies for the purpose of exemplifying the application of this method to the problem of branching. The species belonging to the α-sequence $^{228}_{90}$Rd Th $\xrightarrow{4\alpha}$ $^{212}_{82}$Th B have isotopic number 48. The hypothetical species belonging to the "neighbouring" sequences $^{228}_{89}$Ms Th$_2$ $\xrightarrow{4\alpha}$ $^{212}_{81}$Tl and $^{228}_{91}$Pa $\xrightarrow{4\alpha}$ $^{212}_{83}$Th C similarly have $A-2Z=50$ and $A-2Z=46$, respectively, and the E/Z curves for these values of the isotopic number are provided by our knowledge of the α-disintegration energies of the bodies of the main sequence of the uranium series ($A-2Z=50$), and of the "isolated" α-active bodies $^{212}_{83}$Th C and $^{214}_{84}$Ra C′ ($A-2Z=46$). We have, therefore, for the three "isobaric" sequences ($A=228$) $\xrightarrow{4\alpha}$ ($A=212$), values of individual α-disintegration energies, for one sequence by direct observation, and for the other two sequences by prediction from the E/Z curves. We have, also, direct experimental knowledge of the total β-disintegration energies for the transformations $^{228}_{89}$Ms Th$_2$ $\xrightarrow{\beta}$ $^{228}_{90}$Rd Th and $^{212}_{82}$Th B $\xrightarrow{\beta}$ $^{212}_{83}$Th C. Applying conservation rules we obtain from these data, values of the β-disintegration energies for all species under consideration for which $A-2Z$ is either 48 (the thorium series species) or 50—and the β-disintegration energies of the latter species are essentially values of the electron-capture energy release (with sign reversed) for the former (assuming the rest mass of the neutrino to be zero). We have, in fact, in this way, full information concerning the stability of the five thorium series species concerned, both in respect of β- and capture-disintegration. In the particular case considered the result is that all five species, $^{228}_{90}$Rd Th, $^{224}_{88}$Th X, $^{220}_{86}$Tn, $^{216}_{84}$Th A and $^{212}_{82}$Th B, are capture-stable, and that, with the exception of $^{212}_{82}$Th B, all are β-stable also (in this result $^{216}_{84}$Th A appears β-stable by 0·2 Mev.).

It will be realized that the result just quoted establishes as "pure α-active species" four nuclei of isotopic number 48, namely ($^{228}_{90}$), ($^{224}_{88}$), ($^{220}_{86}$) and ($^{216}_{84}$). Application of the same method to the other radioactive series enables us to extend this list, and finally, after further discussion of the problem, we are led to the following table of such species in the range of Z from 84 to 96 * (Feather, 1947; *see also* Seaborg, 1946).

Pure α-active species ($84 \leqslant Z \leqslant 96$)

Z	A
84	210, 211, 212, (213), 214, 216.
85	215, (217).
86	216, (218), (219), 220, 222.
87	221.
88	(220), (222), 223. 224, 226.
89	225.
90	(224), (226), 227, 228, 229, 230, 232.
91	231.
92	(230), 232, 233, 234, 235, 236, (238).
93	237.
94	(236), 238, 239, 240, 242.
95	241, (243).
96	(240), (242), (243), 244, 245, 246, (248).

* Brackets indicate cases of doubt.

A very obvious fact emerges from the table as given. It is at once clear that the systematics of the pure α-active species included in the table are precisely the same as the systematics of the stable nuclei (for which $Z \leqslant 83$): there is the same evident distinction between even and odd charge numbers, and the same lack of neighbouring isobars, as with the stable elements (*see also* Turner, 1940 a, 1945). Obviously, the pure α-active species of $Z > 83$ are the counterparts, in all essential characteristics, of the truly stable species of smaller Z. Only if we were to discover, on investigating elements of atomic number greater than 96, that none of the isotopes could be identified as a pure α-active species, should we be forced to conclude that we had reached the natural limit of complexity of the nucleus—that we at last knew the number of the elements.

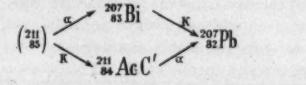
One further reflection is an obvious one in respect of the table just given. It is that the seven species for which α/β branching may be regarded as established are not by any means the only species amongst the heavy radioelements for which α/β branching is energetically possible. It is worth while, perhaps, setting out the reason for the failure to observe the effect in other cases. It derives mainly from the different degrees of dependence which characterize the variation of disintegration probability with disintegration energy in α- and β-disintegration, respectively. This variation being so much more rapid in α-disintegration than it is in β-disintegration (otherwise expressed "very short lived" β-bodies being entirely unknown and "very long lived" β-bodies very nearly so) we have the result that unless the energy of α-disintegration lies within fairly narrow limits (say between 5 and 7 Mev. for elements of $Z \sim 85$) the α-disintegration probability is either so small that no β-disintegration which is energetically possible is at all likely to be characterized by a disintegration probability which is anything like so small, or else it is so large that no β-disintegration, however high the disintegration energy might be, could possibly compete with it as an alternative disintegration mode. Active species may thus appear as pure β-emitters, even though they are α-unstable to the extent of 4 Mev. or more, or as pure α-emitters, if the energy of α-disintegration is greater than, say, 7·5 Mev., although there might be several million electron-volts energy available for β-disintegration. A survey of active species at present known indicates that a rare α-mode might be expected with $^{223}_{87}$Ac K and $^{225}_{88}$Ra (branching ratios approximately 2/1000 and 7/1000, respectively) but suggests that, beyond these two, other known β-active bodies are unlikely to show α/β branching beyond the limiting ratio of $1/10^6$. Similarly, amongst known α-active species the α/β ratio can hardly be less than $10^7/1{\cdot}6$ for $^{215}_{84}$Ac A, although the energy available for β-disintegration is 0·6 Mev.*

§ 3. α/CAPTURE BRANCHING

The only clear case of α/capture branching hitherto reported in the literature is that of the species ($^{211}_{85}$) formed by the (α, 2n) reaction with bismuth (Corson, MacKenzie and Segrè, 1940). The α/capture branching ratio for this body of

* Karlik and Bernert (1944) have reported an observation which might be explained if the α/β branching ratio for AcA were $10^6/5$. Further investigations will be necessary, however, before this interpretation can be regarded as established.

7·5 h. half-value period may be determined directly by α-particle counting, as is evident from the following scheme:



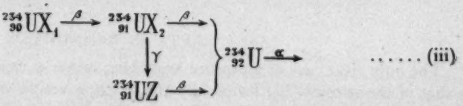
..... (ii)

The value obtained in the original investigation is 40/60. A point of interest attaching to this case of branching is that Ac C′, which is a rare product in the natural actinium series, is here found in the more probable branch. It is likely than an accurate determination of the half-value period of this body (predicted value $2{\cdot}5 \times 10^{-3}$ sec.) will first be made by a study of the K-radiation—α-particle coincidences obtained with a source of $(^{211}_{85})$.

The reason for the non-appearance of capture-active species amongst the classical series of radioelements is, of course, well known. The already considerable neutron excess of the heaviest nuclei is further exaggerated as a result of sequences of α-disintegrations, and the incidence of β-disintegration as a corrective process is a natural result. But, with the production of isotopes of smaller mass number "artificially", capture-active species have been found even with the transuranics (Seaborg, 1946), and since these species are almost certainly α-unstable the possibility of the early discovery of further cases of α/capture branching is evidently open. One example of such a case will suffice. The half-value period of the capture-active $^{235}_{93}$Np has been given by Seaborg as about eight months. From our general knowledge of α-disintegration energies and lifetimes we may hazard the guess that the α/capture ratio for this species will be found to be of the order of $1/10^4$.

§ 4. β/β AND β/γ BRANCHING

The classical case of supposed β/β branching, that of uranium X_1, has been thoroughly studied since the validity of the explanation in terms of nuclear isomerism was securely established by Walling (1931). During the last ten years successive contributions towards final elucidation have been made by Feather and Bretscher (1938), Feather and Dunworth (1938) and by Bradt and Scherrer (1945). As a result it is now almost certain that true β/β branching does not in fact contribute appreciably to the appearance of the two isomers in this case. Feather and Bretscher pointed out that the lifetime of the metastable state required for the explanatory scheme would be the shortest possible if it were assumed that the upper state of the nucleus $(^{234}_{91})$ belongs to uranium X_2 and the ground state to uranium Z, when the scheme would be essentially as given below.

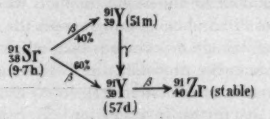


..... (iii)

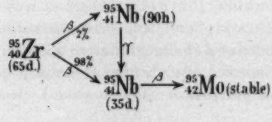
Seven years later Bradt and Scherrer obtained direct evidence for a highly internally converted γ ray of low energy (0·394 Mev.), with an intensity such as would be expected from (iii) and an energy such as would satisfy the scheme within the limits of accuracy of the determinations (maximum β-particle energy

$UX_2 = 2{\cdot}32$ Mev.; maximum β-particle energy $UZ = 0{\cdot}45$ Mev., γ-rays UZ two quanta each $\sim 0{\cdot}8$ Mev. energy *). We have to admit, then, that in this case the branching involved is not the β/β branching of uranium X_1, forming uranium X_2 and uranium Z in alternative modes, but the β/γ branching of uranium X_2, leading to uranium II and uranium Z as immediate products. If we look for the reason why it has taken so long to establish this conclusion by direct experiment, it clearly lies in the short lifetime (1·14 min.) of the most important product, uranium X_2, and in the large value of the β/γ branching ratio (650/1) by which it is characterized.

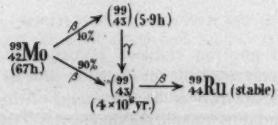
If the classical case is now differently interpreted, other instances have recently been observed for which interpretation in terms of β/β branching is clearly required. Some of the best established of these are represented by the disintegration schemes (iv)–(viii) which follow (Siegel, 1946).



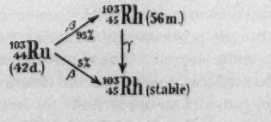
. . . . . (iv)



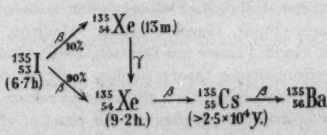
. . . . . (v)



. . . . . (vi)



. . . . . (vii)

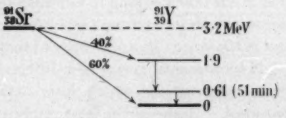


. . . . . (viii)

We may discuss the first of these schemes more fully, from the standpoint of energy balance. The absorption end-points for the β-particles of the 9·7 h. $^{91}_{38}Sr$ have been determined, and limiting energies for the 40 per cent and 60 per cent

* Maximum β-particle energies refer in each case to the most prominent component of the β-particle spectrum.

partial spectra deduced as 1·3 Mev. and 3·2 Mev., respectively. A quantum radiation of roughly 1·3 Mev. energy has also been established. The γ-radiation from the 51 min. metastable state of $^{91}_{39}$Y has been observed, and the energy-value fixed as 0·61 Mev. by absorption measurements. Obviously conservation can be satisfied according to the following level diagram:



For the other cases represented by disintegration schemes (v) to (viii) above, results reported to date appear to be equally satisfactory and no particular feature of the disintegrations merits separate notice—except, incidentally, the very long lifetimes of the product species $(^{99}_{43})$ and $^{135}_{55}$Cs.

As a result of the work done on the fission products we may conclude that the number of certain cases of β/β branching now exceeds the number of established instances of β/γ branching, though it is obvious that, when any β-active nucleus is formed in a metastable state, the probability of β-disintegration from this state (competing with the isomeric de-excitation transition) can never be zero. It is almost certain, therefore, that further investigation will reveal many cases in which this type of branching occurs, for, in fact, there is well-nigh complete formal parallelism between the two phenomena—the formation of nuclear isomers directly in alternative modes of β-disintegration of a single parent species and the formation of a product species equally through the β-disintegration of one or other of a pair of isomeric parents.

§ 5. β^+/CAPTURE BRANCHING

Energetically, when positron emission occurs, capture transformation is always possible: the energy available for the latter process is greater than for the former by $2m_0c^2 - W$, where m_0c^2 is the rest energy of the electron and W is the ionization energy of the product atom in the electron level from which capture takes place. Capture transformation may occur with species which are β^+-stable (as with $^7_4$Be, for example), but all positron-active species must, in the limit, show capture/β^+ branching to a finite degree. The experimental determination of the branching ratio, when the available energy and the characteristics of the nuclear states involved are known, provides an opportunity for testing current theories of β-disintegration, and detailed studies have already been made of the branching ratio in a number of cases (Bradt, Gugelot, Huber, Medicus, Preiswerk and Scherrer, 1945 a, b and c; Good, Peaslee and Deutsch, 1946; Good and Peacock, 1946). All variants of the neutrino theory predict a β^+/capture ratio which decreases with increasing nuclear charge, for a given energy release, but the energy dependence of the ratio is markedly greater on the basis of the Konopinski-Uhlenbeck variant of the theory than on the basis of any type of Fermi theory. For allowed transitions all types of Fermi theory lead to essentially the same value for the ratio.

Different experimental methods have been used for the determination of the branching ratio in different cases. For the 6·7 h. $^{100}_{48}$Cd, Bradt and collaborators (1945 c) have determined the ratio of the number of positrons to the number of

conversion electrons emitted as a result of the 99 per cent internally converted radiation from the 40 sec. metastable state of $^{109}_{47}$Ag which is formed both by positron emission and by K capture, for the isotopic species $^{61}_{29}$Cu and $^{64}_{29}$Cu the same authors (1945 a) have compared the ratios of effects due to annihilation radiation and K x-radiation in an absorption experiment (thus determining with considerable precision the ratio of the branching ratios for these two species, an experimental datum which on any theory must depend almost exclusively on the disintegration energies involved), and for $^{22}_{11}$Na, $^{48}_{23}$V, $^{52}_{25}$Mn, $^{58}_{27}$Co Good, Peaslee and Deutsch (1946) have used a coincidence counting technique. The last-mentioned experiment was possible because in each of the cases studied the product nucleus is left in the same excited state as a result both of positron emission and capture transformation, thus the investigation reduces to a determination of the fraction of disintegrations in which there is correlation in time between the γ-radiation (which is emitted in all disintegrations) and the annihilation radiation which follows the emission of positrons with negligible delay. The same method has been applied by Good and Peacock (1946) in a determination of the β^+/capture branching of $^{65}_{30}$Zn.

The general result of all these investigations is that the original Konopinski-Uhlenbeck theory of β-disintegration is entirely ruled out, whilst Fermi-type theories appear to give reasonably correct "theoretical" results. However, since the cases studied so far are almost all examples of "allowed" transitions, there is little evidence at present to indicate which variant of the Fermi theory is to be preferred. Obviously a more detailed study of this type of branching is well worth while.

§ 6. β/CAPTURE AND β/β^+ BRANCHING

The well-known generalization that, amongst the stable species, isobaric pairs almost invariably differ in atomic number by two units (rather than by one unit) has long been regarded as support for the assumption that the rest mass of the neutrino is effectively zero. If the neutrino mass were identically equal to zero, and if electron capture from the outermost atomic levels could be taken into consideration, we should have to conclude that pairs of "neighbouring" stable isobars (i.e. isobaric species for which $\Delta Z = 1$) were virtually impossible: one or other of the isobaric species would necessarily be unstable (the species of lower Z in respect of β-disintegration or that of higher Z in respect of capture transformation). By the same argument—and on the same assumptions—if $\binom{A}{Z-1}$ and $\binom{A}{Z+1}$ are stable isobars for which $\Delta Z = 2$, clearly the species $\binom{A}{Z}$ must exhibit β/capture branching with finite branching ratio. Because of the further generalization that both A and Z are even for all species paired as stable isobars with $\Delta Z = 2$, we evidently have that, for our "intermediate" (β/capture)-branching species $\binom{A}{Z}$, A must be even and Z odd—thus this particular type of branching is predicted for some sixty or so (A even, Z odd) species which go to form isobaric triads with the known stable species belonging to the pairs.

A survey of lists of artificially-produced active species compiled in 1944 shows that about 45 of these 60 "intermediate" species had then been identified, but it also shows that β/capture (or β/β^+) branching had not been recognized as occurring with more than ten of them. A great deal of careful work is evidently called for in this field.

Two cases of established β/capture branching may be mentioned for their particular interest. In 1937 Weizsäcker suggested that the relatively high abundance of $^{40}_{18}$A on the earth ($^{36}_{18}$A would be expected to be the most abundant argon isotope, which it is not) might be due to its production in the capture transformation of $^{40}_{19}$K. Until then the natural radioactivity of potassium had been regarded as pure β-activity (producing $^{40}_{20}$Ca), but recently Thompson and Rowlands (1943) have obtained evidence for the correctness of Weizsäcker's suggestion. Since the half-value period is of the order of 10^{10} years, this case is exceptional in that both β-disintegration and capture transformation must be highly "forbidden".* The second case is that of $^{64}_{29}$Cu. It was for this well-known active species, of half-value period 12·8 h., that β/β^+ branching was first discovered (Van Voohris, 1936). Now, as already mentioned, it is known that $^{64}_{29}$Cu is subject to capture transformation also. The relative probabilities of β-disintegration, β^+-disintegration and capture transformation are approximately 1·6 : 1 : 2·2.

REFERENCES

Albrecht, 1919. *Sitzungsber. Akad. Wiss. Wien*, IIa, **128**, 925.
Berthelot, 1942. *J. Phys.*, **3**, 17; **3**, 52.
Bittencourt and Goldhaber, 1946. *Phys. Rev.*, **70**, 780.
Bleuler and Gabriel, 1947. *Helv. Phys. Acta*, **20**, 67.
Bradt, Gugelot, Huber, Medicus, Preiswerk and Scherrer, 1945 a. *Helv. Phys. Acta*, **18**, 252 ; 1945 b. *Ibid.*, **18**, 260 ; 1945 c. *Ibid.*, **18**, 351.
Bradt and Scherrer, 1945. *Helv. Phys. Acta*, **18**, 405.
Broda and Feather, 1947. *Proc. Roy. Soc.*, A, in course of publication. (Atomic Energy Report BR 615 (1945).)
Bronson, 1908. *Phil. Mag.*, **16**, 291.
Carlson and Oppenheimer, 1931. *Phys. Rev.*, **38**, 1787.
Chalfin, 1929. *Z. Phys.*, **53**, 130.
Corson, MacKenzie and Segrè, 1940. *Phys. Rev.*, **58**, 672.
De Benedetti and McGowan, 1946. *Phys. Rev.*, **70**, 569; 1947. *Phys. Rev.*, **71**, 380.
English, 1945. Unpublished (Atomic energy report MC 145 (1945)).
Fajans, 1911. *Phys. Z.*, **12**, 369 ; 1912. *Ibid.*, **13**, 699 ; 1926. *Naturwissenschaften*, **14**, 963.
Fajans and Voigt, 1941. *Phys. Rev.*, **60**, 619.
Feather, 1929. *Phys. Rev.*, **34**, 1558 ; 1945. Unpublished (Atomic Energy Report BR 640 (1945)) ; 1946. *Proc. Roy. Soc., Edin.*, **62**, 211 ; 1947. *Nuclear Stability Rules* (Cambridge University Press, in preparation).
Feather and Bretscher, 1938. *Proc. Roy. Soc.*, A, **165**, 530.
Feather and Dunworth, 1938. *Proc. Roy. Soc.*, A, **168**, 566.
Flammersfeld, 1947. *Z. Naturforsch.*, **2a**, 85
Flügge and Krebs, 1944. *Naturwissenschaften*, **32**, 71.
Good and Peacock, 1946. *Phys. Rev.*, **69**, 680.
Good, Peaslee and Deutsch, 1946. *Phys. Rev.*, **69**, 313.
Guében, 1933. *Ann. Soc. Sci., Bruxelles*, **53**, 115.

* The observations of Thompson and Rowlands have since been confirmed by Bleuler and Gabriel (1947), and Flammersfeld (1947) has reported that the naturally occurring $^{176}_{71}$Lu also shows similar branching, each mode being highly forbidden.

HAHN, 1906. *Phil. Mag.*, **11**, 793 ; 1921. *Ber. dtsch. chem. Ges.*, B, **54**, 1131.
HAHN and ROTHENBACH, 1913. *Phys. Z.*, **14**, 429.
HENDERSON, 1934. *Proc. Roy. Soc.*, A, **147**, 572.
JACOBSEN, 1927. *Nature, Lond.*, **120**, 874.
JOLIOT, 1930. *C.R. Acad. Sci., Paris*, **191**, 132.
KARLIK and BERNERT, 1943. *Naturwissenschaften*, **31**, 298 ; **31**, 492; 1944. *Z. Phys.*, **123**, 51.
KOVARIK and ADAMS, 1938. *Phys. Rev.*, **54**, 413.
KURTCHATOW, KURTCHATOW, MYSSOWSKY and ROUSSINOW, 1935. *C.R. Acad. Sci., Paris*, **200**, 1201.
MARSDEN and DARWIN, 1912. *Proc. Roy. Soc.*, A, **87**, 17.
MARSDEN and PERKINS, 1914. *Phil. Mag.*, **27**, 690.
MARSDEN and WILSON, 1913. *Phil. Mag.*, **26**, 354.
MERCIER, 1926. *C.R. Acad. Sci., Paris*, **183**, 962.
MEYER, HESS and PANETH, 1914. *Sitzungsber. Akad. Wiss. Wien*, IIa, **123**, 1459.
PEREY, 1939. *C.R. Acad. Sci., Paris*, **208**, 97.
RONA and SCHINTLMEISTER, 1938. *Sitzungsber. Akad. Wiss. Wien*, IIa, **147**, 49.
ROSENBLUM, 1929. *C.R. Acad. Sci., Paris*, **188**, 1401.
RUTHERFORD, LEWIS and BOWDEN, 1933. *Proc. Roy. Soc.*, A, **142**, 347.
RUTHERFORD, WARD and WYNN-WILLIAMS, 1930. *Proc. Roy. Soc.*, A, **129**, 211.
RUTHERFORD, WYNN-WILLIAMS and LEWIS, 1931. *Proc. Roy. Soc.*, A, **133**, 351.
RUTHERFORD, WYNN-WILLIAMS, LEWIS and BOWDEN, 1933. *Proc. Roy. Soc.*, A, **139**, 617.
SCHINTLMEISTER, 1938. *Österr. Chem. Z.*, **41**, 315.
SEABORG, 1946. *Science*, **104**, 379.
SEGRÈ, 1946. Isotope chart (circulated privately).
SIEGEL, 1946. *J. Amer. Chem. Soc.*, **68**, 2411. (Plutonium Project Report.)
SODDY, 1909. *Phil. Mag.*, **18**, 739 ; 1917. *Proc. Roy. Inst.*, **22**, 117.
SZILARD and CHALMERS, 1935. *Nature, Lond.*, **135**, 98.
THOMPSON and ROWLANDS, 1943. *Nature, Lond.*, **152**, 103.
TURNER, 1940. *Phys. Rev.*, **57**, 950 ; **58**, 181; 1945. *Rev. Mod. Phys.*, **7**, 292.
VAN VOORHIS, 1936. *Phys. Rev.*, **49**, 876.
WALLING, 1931. *Z. phys. Chem.*, B, **14**, 290 ; 1932. *Z. Phys.*, **75**, 425.
WEIZSÄCKER, 1936. *Naturwissenschaften*, **24**, 813 ; 1937. *Phys. Z.*, **38**, 623.
WRIGHT, 1932. *Proc. Camb. Phil. Soc.*, **28**, 128.

THE NEUTRINO AND THE RECOIL OF NUCLEI IN BETA DISINTEGRATIONS

By B. PONTECORVO,
National Research Council, Chalk River, Ontario

§1. INTRODUCTION

A GREAT deal of work on the recoil of nuclei in radioactive transformations has been done in the development of classical radioactivity. Most of this work, however, referred to α disintegrations; because the mass of an α particle is much higher than the mass of a β particle, the energy of the recoil nucleus in α disintegrations is several orders of magnitudes higher than the energy of the recoil nucleus in β disintegrations. For example, when radium A disintegrates by emission of an α particle, the recoil nucleus radium B acquires an energy of 110,000 ev., equal to about 2 % of the energy of the α particle. When radium B disintegrates by emission of a β particle, the recoil nucleus radium C acquires only an energy of 0·4 ev.

In 1909, Hahn and Meitner (1909) introduced the recoil method of radioactive analysis. They found that a negatively charged plate, placed near a plate coated with radioactive elements, became radioactive, by collecting the recoil atoms, and this is today the basis for the preparation of active deposit sources. The method has since been used with great success in preparing actinium C″ and thorium C″ in a pure state from actinium C and thorium C, and in many other researches, which are described in textbooks of classical radioactivity (e.g. Rutherford, Chadwick and Ellis, 1930). In addition the properties of the recoil atoms from α and β disintegrations were studied extensively. Some of these properties, such as range and charge of α recoils are reasonably well established. As for the recoil atoms in a β transformation, some conflicting results were obtained, which are described by Rutherford, Chadwick and Ellis (1930). We may summarize by stating that:

(i) The recoil atom, when coming from a coated surface, is usually uncharged.

(ii) The efficiency in the collection of β recoils is much smaller (only a few per cent) than in the case of α recoils, and even in a good vacuum depends greatly on the purity of the radioactive film and the surface properties of its support.

(iii) The velocity of the recoil atoms is so low that they are often reflected from the receiving surface: in fact the efficiency of collection may increase by one order of magnitude when the collecting surface is maintained at a low temperature.

In this *Report*, however, we are not concerned with the properties of recoils due to β disintegrations as such. Rather, we plan to give an account of those aspects of the subject which have a direct bearing on the understanding of the β-ray process itself and on the main assumption of the β-ray theory, the neutrino hypothesis. Experimental work to this end has been carried out fairly recently and it is the aim of this paper to report the progress obtained so far.

§2. THE FUNDAMENTAL ASSUMPTION OF THE β-RAY THEORY

It will be useful to recall briefly the main experimental facts of β disintegration:

(i) In a β transition, the atomic number Z of the nucleus changes by one unit while the mass number remains unchanged. Since the angular momentum of a nucleus is an integer or half integer, according as its mass number is even or odd, a consequence of this is that the difference in the angular moments of the nuclei Z and $Z \pm 1$ must be an integral number of units of $\frac{h}{2\pi}$. This fact cannot be explained if the β process consists of the emission of only one particle, the electron, having spin 1/2.

(ii) The β spectrum is continuous, while the parent and the daughter states correspond to well defined energy values of the nuclei Z and $Z \pm 1$.

(iii) The difference in energy between the initial and final states of the nucleus involved in a β transition is equal to the upper energy limit of the continuous spectrum. Calorimetric measurements (Ellis and Wooster, 1927; Meitner and Orthmann, 1930; Lecoin and Zlotowski, 1939) have shown that the difference between the upper energy and the average energy of the spectrum is not emitted in any form which is absorbed in a calorimeter designed to absorb radiations as penetrating as γ rays.

These facts can be reconciled only with one of the following alternative assumptions:

(*a*) The laws of conservation of energy and of spin do not hold in a single β process.

(*b*) The laws of conservation of energy and of spin are valid, but a new hypothetical particle having zero charge and spin 1/2, *the neutrino*, is emitted together with a β particle, in such a way that the energy available in the β transition is shared between the β particle and the neutrino. In a β transition, consisting of the capture of an orbital electron by a nucleus, the neutrino emitted would have a well-defined energy corresponding to the fact that the bound electron has a discrete energy.

On the basis of the latter assumption, made by Pauli in 1931, Fermi (1934) has developed a consistent and quantitative theory of the β decay which accounts for the essential features of the β process.

The physical properties of the neutrino were assigned to it at the same time as its existence was postulated, in order to explain the fundamental facts mentioned above. These properties are zero charge, spin 1/2, and small magnetic moment. The value of the mass which is zero, or at least small in comparison with the electron mass, was assigned to the hypothetical particle in order to explain the shape of the continuous β-ray spectra on the basis of Fermi's theory, and also is required by other recent evidence (Lyman, 1939).

We shall see later that experiments on recoils of nuclei in β decay can give direct information, in principle, on the mass of the neutrino. Concluding these introductory remarks we can state that Fermi's theory has been very successful in explaining the experimental facts, on the basis of the neutrino hypothesis. The neutrino hypothesis is consequently strongly supported.

The evidence for the existence of the neutrino, however, is very indirect. It was suggested by many physicists that experiments on the recoil of the nuclei in β decay might confirm in a decisive way the existence of neutrinos. As will be shown below, however, experiments of this type can only either *disprove* the neutrino hypothesis or *increase* the "indirect" evidence for the existence of the neutrino (and at the same time give important information on the β decay): a "direct" proof of the neutrino existence, *different in character* from the evidence already available, cannot be obtained by recoil experiments.

§ 3. THE INTERPRETATION OF EXPERIMENTS ON NUCLEAR RECOIL IN β DECAY ON THE BASIS OF THE NEUTRINO HYPOTHESIS

In the considerations which follow immediately we shall admit that the law of conservation of energy is valid, i.e. we shall admit the existence of the neutrino according to the alternative (*b*) mentioned above.

Let us consider a single β disintegration, consisting of the emission of a β particle and of a neutrino. The energy of the β particle E_β under consideration may be measured. The energy of the neutrino E_ν associated with the β particle is then known, provided one knows the energy $E_{\max}$ of the upper limit of the β spectrum: $E_\nu = E_{\max} - E^\beta$. The momentum of the neutrino p_ν is related to its energy by the well known relativistic formula

$$p_\nu^2 c^2 = E_\nu^2 + 2E_\nu \mu_\nu c^2,$$

where μ_ν is the neutrino mass.

The momentum of the neutrino, on the other hand, is known if the momenta of the recoil nucleus and of the β particle are measured, in magnitude and direction, since the resultant of the momenta of the three particles must be zero. We can see then that the value of μ_ν can be obtained, at least in principle, by experiments on recoils of nuclei in β decay. It should be noticed that a constant value of μ_ν should result from observations on a β particle and associated recoil, independently of the energy of the β particles. An experiment of this type, *when performed with sufficient accuracy*, would give the maximum information which can be obtained by investigating recoils of nuclei in β decay. It will be useful here to mention specifically all the information which could be obtained by a sufficiently accurate experiment.

(i) A recoil experiment is capable of demonstrating whether the resultant of the momenta of the β particle and of the recoil nucleus is zero. In the event of this resultant being found to be zero, the neutrino hypothesis should be abandoned, since it would be unsatisfactory to abandon the fundamental law of the conservation of linear momentum in order to hold to the neutrino hypothesis. On the other hand, if the resultant of the momenta of the β particle and of the nuclear recoil is different from zero, the law of the conservation of linear momentum would require that a third particle (the neutrino) participate in the β process. We shall see below that actually the experiments performed point to the non-conservation of linear momentum of the β particle and of the recoil nucleus. In connection with the alternative (*b*), it will be possible to conclude, by these experiments, that not only the conservation of spin and of energy, but also the conservation of linear momentum, requires the existence of the neutrino.

(ii) Recoil experiments, as indicated above, permit in principle the determination of μ_ν. Should the value of μ_ν turn out to vary as different energies of the β particle are involved, this should be a conclusive proof against the neutrino hypothesis. The difficulty in obtaining significant information on the mass problem, however, is so great, that this point has not yet been cleared and it is doubtful whether it will ever be by such experiments.

(iii) Recoil experiments are capable of giving information on the angle between the directions of emission of the neutrino and the β particle. The angular distribution is related in turn to the form of the interaction (Bloch and Møller, 1935; Hebb, 1938) between the heavy particle and the electron-neutrino field. Information obtained on this point until now is only of a qualitative nature; it indicates that the electron and the neutrino are emitted with a positive value of the average cosine of their emission directions.

§4. LEIPUNSKI'S EXPERIMENTS

The first recoil experiment on the neutrino problem was performed by Leipunski (1936). Leipunski's paper is very brief and the description of the apparatus is not given in detail, probably because the writer intended to perform more quantitative experiments at a later time. Leipunski used as a radioactive source C^{11}, a radio-element emitting positrons with a half-period of 21 minutes, produced in the reaction $B^{10}+H^2 \rightarrow C^{11}+n_0^1$. The radioactive carbon was adsorbed in the form of $C^{11}O$, or $C^{11}O_2$ on a flat metal disc cooled by liquid air. The analysis of the recoils was limited to the small fraction of the recoil ions which were positively charged. The positive recoil ions were analysed by means of a retarding field between the cooled disc supporting C^{11} and a grid. The detection, which presents probably the most serious difficulty in experiments with recoils, was effected by using the property of low energy ions of ejecting electrons on striking a metal surface. This property, incidentally, was later used by Allen (1942) in his experiments. The ions having passed the analysing grid mentioned above, were accelerated by 5000 ev. placed between the grid and a metallic plate, on striking which they liberated secondary electrons. These electrons were accelerated and attracted into a thin-window Geiger counter. By measuring the number of discharges of the counter as a function of the analysing retarding voltage X of the grid, Leipunski was able to plot a curve showing the integral distribution in energy of the recoils, i.e. showing the number of recoils having an energy higher than X.

Leipunski's experiment is extremely qualitative, as he himself points out due to the very serious experimental difficulties introducing systematic errors and to the lack of statistical accuracy. However, he was able to compare his experimental curve with the curve which can be calculated from the β spectrum assuming the conservation of the momentum of the recoil nucleus and the associated β particle (no neutrino). The experimental curve shows an energy of recoils considerably greater than would have been expected without neutrinos. Thus Leipunski's experiment for the first time indicated that the law of the conservation of momentum, in addition to the laws of conservation of energy and of spin, can be maintained only if a third particle participates in the process.

The greatest difficulty in improving the accuracy of Leipunski's work would arise from the interpretation of experiments in which the radioactive source is "supported" (as opposed to the experiments in which the radioactive source is a gas).

§ 5. THE EXPERIMENTS OF CRANE AND HALPERN

Crane and Halpern (1938, 1939) used a gaseous compound as a source and a cloud chamber for the observation of the β particles and of the recoils. Cl^{38} was used because of its convenient lifetime of 37 minutes, of the high upper energy of the β spectrum emitted, and of its relatively low atomic weight. The difficulty in such a scheme is that the length of track made by the recoil nucleus is very short even in a low-pressure Wilson chamber, although the recoil of the nucleus is expected to produce a number of ion pairs. Crane and Halpern overcame this difficulty by allowing the ions to diffuse into a cluster several millimetres in diameter before the alcohol vapour was made to condense upon them. In this way they were able to count the number of individual droplets. This was accomplished by removing the clearing electric field a fraction of a second before the expansion. By applying a magnetic field to the chamber, it was possible in this way to measure at the same time the momentum of the electron and the number of droplets produced by the recoil in *each individual β disintegration*.

By using their beautiful cloud chamber technique, Crane and Halpern obtained a few dozen photographs, capable of being analysed, and yielding the momentum of the β particle and the number of droplets produced by the associated recoil. Electrons having energies smaller than 1 Mev. were disregarded in the selection of the pictures for analysis, for the following reason. The Cl^{38} β spectrum consists of two components, one associated with a γ radiation and having a β spectrum of upper energy about 1 Mev., and the other, not accompanied by γ radiation, having a β upper energy of about 5 Mev. Because of their possible association with a γ radiation, the electrons of low energy must be disregarded in an experiment aimed to give information on the neutrino problem.

The relation between the number of droplets and the energy of the recoil atom, unfortunately, is completely unknown. However, an important conclusion is obtained from the experiment, even without making detailed assumptions on the relation between the energy of the recoil and the number of droplets produced. In fact, Crane and Halpern, as illustrated in figure 1, found that recoils producing a large number of droplets are associated with β particles of low momentum as often as they are with β particles of high momentum. This fact shows that the momentum *is not conserved between the electron and the recoil nucleus*.

The experiment would be capable of throwing light on the problem of the angular distribution between electrons and neutrinos, were the relation between the number of droplets and the energy of the recoil known. One would know then the energies of the three particles participating in a single β process, and consequently solve for the angle ϕ between the emission directions of the electron

and of the neutrino. In absence of a better relation,* Crane and Halpern assumed that the energy of the recoil atoms is proportional to the number of droplets produced. In addition, they normalized their energy scale by considering a particular β-decay process, that they were fortunately able to photograph, in which the β particle is emitted with practically the total maximum energy of the spectrum (this process corresponds to the point of momentum 11 mc in the diagram of figure 1). In this case the energy of the recoil atom is known, since the neutrino has no energy and its direction consequently is immaterial. Crane and Halpern were able to obtain in this way the angular distribution of the directions of the neutrino and electron, and compare their observed distribution with the theoretical predictions of Fermi (1934) and of Konopinski and Uhlenbeck

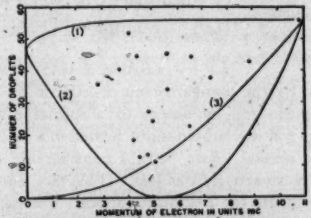


Figure 1. Plot of the number of droplets in the cluster against the momentum of the emitted beta-particle. Each point in the diagram represents an individual disintegration, and the number of droplets in each cluster has been corrected for the number of droplets which is due to the escaping beta-particle, so that the number plotted represents only those due to the recoil atom. The curves shown have been calculated for the two limiting cases, (1) that in which the electron and neutrino escape in the same direction and (2) that in which they escape in opposite directions. Curve (3) shows what we should expect if momentum were conserved between the beta-particle and the nucleus (no neutrino).
(*Reproduced, by kind permission, from the ' Physical Review '.*)

(1935). According to Fermi's form of interaction of the heavy particle with the neutrino field (Bloch and Møller, 1935; Hebb, 1938) the probability that the neutrino will be emitted at an angle ϕ with the direction of the electron within the solid angle $d\Omega$ is $\frac{1}{4\pi}(1+\frac{v}{c}\cos\phi)d\Omega$, whereas this probability is $\frac{1}{4\pi}(1-\frac{v}{c}\cos\phi)d\Omega$ according to the K.U. form of interaction (v is here the velocity of the electron). The experiments tend to show a better agreement with Fermi's predictions. However, the evidence is very slight, in view of the considerable possible errors and the small number of single β processes studied.

§6. THE EXPERIMENTS OF ALLEN

The work of Allen (1942) is characterized by new and interesting features. First, he used a secondary-emission multiplier in order to detect recoils and,

* For a discussion of the problem see Crane and Halpern (1938, 39), and Wertenstein (1938). Crane and Halpern have given evidence that a considerable number of droplets produced by recoils are due to the dissociation of gas molecules, in addition to the ionization.

secondly, he made use of a radioactive source, Be^7, decaying by K capture rather than by β-ray emission (the use of Be^7 in recoil experiments was first suggested by Kang Chang Wang (1942)). Since the K-capture decay is not complicated by emission of electrons, the very existence of a nuclear recoil would be in contradiction with the law of the conservation of the momentum, unless the neutrino participates in the process. It should be noticed that neutrinos emitted in a K-capture process are mono-kinetic, so that the measurement of the energy of the recoil, if accurate enough, would yield directly the mass of the neutrino, provided the energy liberated in the reaction (i.e. almost exactly the neutrino energy) is known.

The reactions involved are

$$Be^7 + K\text{-capture} \begin{cases} \nearrow Li^7 + \text{neutrino} + 0{\cdot}87 \text{ Mev. } (90\%) \\ \searrow (Li^7)^* + \text{neutrino} + 0{\cdot}42 \text{ Mev. } (10\%) \\ \qquad \downarrow \\ \qquad Li^7 + \gamma \text{ radiation of } 0{\cdot}45 \text{ Mev.}; \end{cases}$$

the Be^7 was prepared by the bombardment of lithium by deuterons. Be^7 was deposited on a platinum foil by means of a special evaporation technique. Although the platinum foil was subsequently heated in a vacuum at a temperature of about 1450° C. for several hours, with a considerable amount of platinum evaporation, practically no activity was lost. This fact, showing that the radioactive atoms diffused inside the platinum support, was expected to result in a continuous energy distribution of recoils from the source. This, and the adsorption of a gas layer on the surface of the source, i.e. the impossibility of obtaining a truly thin source from a solid support, makes the interpretation of the experiment quite difficult.

It was found that a considerable gain in the intensity detected by the multiplier was obtained by accelerating the ions before the application of the analysing retarding field. Consequently, the ions were accelerated from the source to a grid by 176 volts and subsequently retarded by 176 volts plus the retarding analysing voltage. By plotting the number of pulses of the multiplier as function of the analysing voltage, it is possible to determine the maximum energy of the recoils. This maximum energy was found to vary according to the "history" of the source-supporting platinum strip. It was found, for example, that the maximum energy of the recoils was 47 ev., when the measurements were made one hour after heating the source, and 41 ev. when the measurements were made two hours after heating the source. By studying in detail this fact it was possible to apply a correction to the results and the maximum energy of the recoils was found to be about 48 ev.

The observed recoils are not due to the emission of the 0·45-Mev. γ radiation, since their value, 48 ev., is definitely higher than the value 15·6 ev. which would correspond to the γ recoils. As a further check Allen showed that there were no coincidences between the recoils and the γ radiation emitted by the source.

It is evident from the very existence of fairly energetic recoils that the law of the conservation of momentum does not hold unless the neutrino is assumed to participate in the K-capture process. As for the value of the recoil energy, about 48 ev., we may compare it with the value of 58 ev., which may be calculated assuming zero mass of the neutrino and 0·87 Mev. for the energy given to the

neutrino. The difference of about 10 ev., according to Allen, is due to the loss of energy of the recoil ions in passing through the gas layer of the surface of the source, and also to the work function of the platinum strip. The accuracy of these experiments is obviously not sufficient to obtain information on the mass of the neutrino, in fact the value of the recoil energy would be lowered only by 1 ev. if the mass of the neutrino were assumed to be 1/5 of the electron mass (this value being the upper limit for the neutrino mass as required by β-spectra evidence) instead of zero.

§7. THE EXPERIMENT OF JACOBSEN AND KOFOED-HANSEN

The method of Jacobsen and Kofoed-Hansen (1945) is essentially different from the previous methods. The Danish physicists pointed out that for the quantitative determination of the recoil energy, it is rather essential that the radioactive source be a *gas at a pressure so low that the mean free path is large compared with the dimensions of the apparatus.* It is thus certain that the recoil atoms will be charged, and their recoil energy can be determined by a retarding field. The gas must also be mono-atomic in order to avoid the possibility of the recoil atom sharing its energy with the other atom or atoms of a molecule.

They observed, in addition, that the problem of the detection would be simplified if the daughter substance formed in the decay (i.e. the recoil nucleus itself) was radioactive, because the radioactivity collected as a function of a retarding field would provide an easy method of measuring the recoil energy.

Among the noble gases produced in uranium fission, it is to be expected that some will fulfil the above requirements, since fission fragments give rise to series of radioactive elements. In fact, the Danish physicists used Kr^{88}, a radio-element having a lifetime of about 2·7 hours produced in uranium fission, as a gaseous source introduced into an apparatus kept at a pressure of 10^{-3} to 10^{-4} mm. Hg. The daughter substance, Rb^{88}, is radioactive with a period of about 18 minutes. By selecting convenient times of irradiation, they were able to collect Rb^{88} practically pure as active deposit. The principle of the experiment is shown in figure 2.

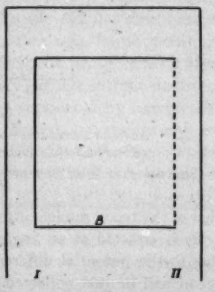


Figure 2

(Reproduced, by kind permission, from the 'Matematisk-Fysiske Meddelelser af Det Kongelig Danske Videnskabernes Selskab'.)

A square metal box B, with one end consisting of a wire gauze was introduced into a vessel containing the radio-element Kr^{88}, produced by fission. The two metal plates I and II are kept at a positive potential X relative to the box and are at equal distances from the wire gauze and the opposite end of the box. After Kr^{88} was kept in the box for a convenient time, the activities (due to Rb^{88}) of plates I and II were measured. The difference between the activities of I and II is due to those recoils which start from the interior of the box and are capable of overcoming the potential difference X. By varying X, it is then possible to

measure the energy distribution (integral) of the recoils, since the difference in activities of the two plates I and II is proportional to the number of recoil atoms with energy higher than X. In figure 3 *a*, the activities of the two collecting plates (actually the activities of two aluminium foils attached to the inner side of plates I and II were measured) are plotted as a function of the retarding potential. Figure 3 *b* shows the difference between the curves in figure 3 *a*, the difference at zero voltage being taken as unity. The ordinate of figure 3 *b* is then the fraction of recoil atoms, starting from the interior of the box, which have sufficient energy to overcome the potential difference X.

It will be useful now to discuss the characteristics of the method. While the advantages of the method are considerable, some difficulties remain, as pointed out by the authors. Thus the method does not give the energy of the recoil in a single β process together with the energy of the associated β particle; it permits only the comparison of the energy distribution of the β particles with

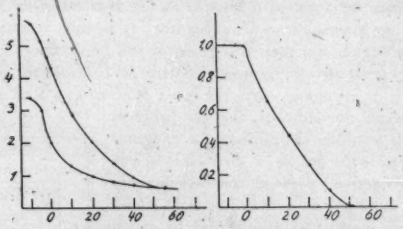


Figure 3 *a*. Figure 3 *b*.

Abscissa: retarding potential X in volts. Ordinate: Figure 3 *a*, activities of collecting foils. Figure 3 *b*, fraction of recoil atoms with energy greater than X.

(Reproduced by kind permission, from the 'Matematisk-Fysiske Meddelelser af Det Kongelige Danske Videnskabernes Selskab'.)

that of the recoil nuclei. Again, let us suppose that a recoil of energy ϵX electron-volts is emitted at an angle ϕ with the direction of the retarding electric field due to the potential difference X volts between the gauze and the plate. The recoil will be just collected only if ϕ is zero, while all the recoils emitted at an inclination $\phi > 0$ with the electric field will not be collected. This results in a very high correction which should be applied to the curves of figure 3 *b*, The authors, however, in discussing in detail this point, arrived at the conclusion that it is not worth while to apply this large correction because of the uncertainties affecting its evaluation. Rather it is sufficient to keep in mind that:

(*a*) The geometrical effect in question will not affect the *upper limit* for the energy of the recoil atoms as determined in curves such as that of figure 3 *b*; and

(*b*) the *true* integral energy distribution of the recoil atoms will be given by a curve which, in figure 3 *b*, would lie anywhere *above* the measured points. With these points in mind it is possible to interpret the experiments as follows.

First, the observed maximum energy of the recoils is 51·5 ev., with an accuracy of about 1%, in very good agreement with the value of about 50 ev. which can be calculated from the approximate knowledge of the maximum energy (2·4 Mev.) of the β spectrum. From this agreement one can deduce the fact, which has

usually been assumed without much experimental evidence, that, apart from the rest mass of the electron and of the neutrino, the energy released in a β disintegration is determined by the upper limit of the β spectrum.

Second, by comparing the "observed" integral distribution in energy of the recoils (curve of figure 3*b*) with the curve giving the distribution of the recoil energies which might be expected if only electrons were emitted, one finds that the actual number of recoil atoms with energies near the upper limit is much larger than can be accounted for, assuming the laws of conservation of energy and momentum and emission of electrons only (no neutrino).

Third, once the existence of the neutrino is assumed, some deduction may be made regarding the angular distribution of the neutrino and electron. The observed distribution may be compared with the distribution which can be expected theoretically by assuming various angular distribution laws and consequently various forms of interaction between the heavy particle and the neutrino field. It is found that the neutrino and electron tend to go in the same direction, although the accuracy is certainly not sufficient to determine actually the angular distribution law.

§8. CONCLUSION

We can summarize by stating first that all the experiments performed until now have shown that the momentum is not conserved in a β disintegration, unless the neutrino participates in the process. As for the angular distribution of the electron and neutrino directions, some indication was found pointing to a tendency for the two particles to go in the same direction, in agreement with the original Fermi form of interaction between the heavy particle and the neutrino field; although this evidence is not very strong, it is supported by completely independent evidence (e.g. Konopinski (1943)) obtained by the study of β spectra, according to which Fermi's interaction is preferable to the Konopinski and Uhlenbeck interaction. This most interesting point will probably be solved in the future with definite results by recoil experiments, which are a very powerful tool to investigate this particular problem.

As for the mass of the neutrino, no value was obtained by recoil experiments and it is doubtful whether future experiments will be able to yield information on this point.

In future developments of recoil experiments, the most promising techniques are the use of secondary-emission multipliers as detectors of recoils and the use of a monatomic gas at low pressure as a radioactive source. Among the radioactive sources of this type, two seem to be particularly interesting: one is He^6, and the other is the "neutron gas". The neutron is expected to be radioactive with a period of about 20 minutes; the very intense beams of thermal neutrons provided by chain-reacting piles will probably give the possibility of studying the characteristics (angular distribution of the proton and electron emitted etc.) of this fundamental β process, consisting of the transformation of a neutron into a proton.

Direct evidence for the existence of the neutrino, i.e. evidence different in character from that which arises, for example, from the existence of continuous β spectra, cannot be obtained by experiments on recoils. The only way of obtaining *direct evidence* for the neutrino's existence is by the detection of some

specific process produced by *free neutrinos*, i.e. a process produced by neutrinos after they have been emitted in a β disintegration. Inverse β transformations produced by neutrinos are processes of this type and certainly are produced by neutrinos, if neutrinos exist at all. They consist of the simultaneous absorption of a neutrino and emission of a β particle. It is clear, on thermodynamical grounds, that such processes must have an extremely low yield, since their inverse, the β process, is so unlikely. It has been currently stated in the literature (e.g. Bethe and Peierls, 1934), that, due to the low cross-section ($\simeq 10^{-42}$ cm^2), an inverse β process produced by neutrinos cannot be observed. This statement, however, seems to be too drastic. A method which might make it possible to observe β processes produced by neutrinos has been suggested by the author (1946).

REFERENCES

Allen, J. S., 1942. *Phys. Rev.*, **61**, 693.
Bethe, H. A. and Peierls, R., 1934. *Nature, Lond.*, **133**, 532.
Bloch, F. and Møller, C., 1935. *Nature, Lond.*, **136**, 911.
Crane, H. R. and Halpern, J., 1938. *Phys. Rev.*, **53**, 789.
Crane, H. R. and Halpern, J., 1939. *Phys. Rev.*, **56**, 232.
Ellis, C. D. and Wooster, W. A., 1927. *Proc. Roy. Soc.*, A, **117**, 109.
Fermi, E., 1934. *Z. Phys.*, **88**, 161.
Hahn, O. and Meitner, L., 1909. *Z. Phys.*, **10**, 697.
Hebb, M. H., 1938. *Physica*, **5**, 701.
Jacobsen, J. C. and Kofoed-Hansen, O., 1945. *Kgl. Danske Vid. Selsk. Mat.-Fys. Medd.*, **23**, no. 12.
Kang Chang Wang, 1942. *Phys. Rev.*, **61**, 97.
Konopinski, E. J., 1943. *Rev. Mod. Phys.*, **15**, 209.
Konopinski, E. J. and Uhlenbeck, G. E., 1935. *Phys. Rev.*, **48**, 7.
Lecoin, M. and Zlotowski, I., 1939. *Nature, Lond.*, **144**, 440.
Leipunski, A. I., 1936. *Proc. Camb. Phil. Soc.*, **32**, 301.
Lyman, E. M., 1939. *Phys. Rev.*, **55**, 234.
Meitner, L. and Orthmann, W., 1930. *Z. Phys.*, **60**, 143.
Pontecorvo, B., 1946. Nuclear Physics Conference, Montreal, Sept.
Rutherford, E., Chadwick, J. and Ellis, C. D., 1930. *Radiations from Radioactive Substances* (Cambridge : University Press).
Wertenstein, L., 1938. *Phys. Rev.*, **54**, 306.

FERROMAGNETISM

By EDMUND C. STONER, F.R.S.,
Department of Physics, The University, Leeds

§1. INTRODUCTION

THE magnetic and associated properties of substances which behave magnetically like iron have long provided a happy hunting ground if not for the "smoother pebble than ordinary" at least for the more curiously shaped one. The rich diversity of ferromagnetic phenomena; the perennial challenge to skill in experiment and to physical insight in coordinating the results, the vast range of actual and possible applications of ferromagnetic materials, and the fundamental character of the essential theoretical problems raised have all combined to give ferromagnetism a width of interest which contrasts strongly with the apparent narrowness of its subject matter, namely certain particular properties of a very limited number of substances.

With so many incentives to the investigation of ferromagnetism in its various aspects, the published work is very extensive, and in any review in moderate compass of even the more recent developments, it is inevitable, if something other than an annotated list of references is to be presented, that much of the work must be passed over without specific mention even though it is of interest in itself, and may have played a part in contributing to the general outlook that has been reached. Moreover, it seems appropriate in the present article to consider work published over a longer period than is usual in the *Reports on Progress in Physics*, and not only to make specific reference to papers published as far back as 1934, but also, since there have been no previous articles concerned exclusively with ferromagnetism in these *Reports*, to give a relatively full account of the background provided by the earlier work. It is hoped that this will prepare the ground for possible reports in the future dealing in greater detail with special aspects of the subject.

During the five years before the war rapid progress was being made in the experimental investigation and theoretical understanding of ferromagnetism. There are good surveys of some of the work of this period in such secondary sources as the books and review articles which are listed in the references, but it is not easy to pick up the threads of the developments of the immediate pre-war years and to proceed immediately from the point that had then been reached. The situation is in some respects similar to that after the previous world war, and it is perhaps worthy of remark that the effectively stationary, or even retrogressive, period in the fundamental development of many branches of science was much longer than the actual duration of that war; in magnetism, for example, it was not until well on in the twenties that much of the work of the period 1909 to 1914, and even earlier, became generally known and appreciated as a basis for further development. In a survey of ferromagnetism at the present time, therefore, it seems much more important to devote adequate attention to developments up to 1939 in connection with central problems than to include reference to all the published papers on ferromagnetism since that date; which last, indeed, alone, would give a very ill-balanced impression of the state of the subject.

The central themes in ferromagnetism are those connected with the conditions

for the occurrence of ferromagnetism, with the magnitude of the saturation magnetization and its temperature variation in different ferromagnetics, and with the variation of magnetization with field. Linking the first with the second there are two major themes, and it is with these that this survey will be mainly concerned. Although the particular phenomena under consideration are peculiar to ferromagnetics, and their investigation involves experimental methods which are sometimes rather specialized, the theory should be looked upon as part of the general theory of the solid state; progress in the theory of ferromagnetism has in fact been characterized by the gradual elimination of special *ad hoc* hypotheses, which, invaluable though some of them have been, should be regarded as temporary scaffolding rather than part of the permanent structure. The first and more fundamental theme involves the consideration of the interactions between atoms, and the statistical treatment of large numbers of atoms as aggregated in a solid; there are links with the chemical theory of valency, though it is the metallic type of binding, which is not typically "chemical", which is of predominant importance. The second theme is bound up with the general problem of structure-sensitive properties which, although often among the most obvious and technologically important properties of metals and alloys, are quite inexplicable in terms of perfect crystal lattices, and indeed, apart from the near certainty that they are associated with structural irregularities, largely unaccounted for in any terms. The variation with field of the magnetization of a ferromagnetic is markedly dependent on structural irregularities of a kind which may equally occur in non-ferromagnetic materials, so the investigation of the low field behaviour is not only interesting and important in itself but is also potentially contributory to the elucidation of a much wider problem.

The subject matter which might most appropriately be considered under the title "Ferromagnetism" has been thought of as falling into three parts: the first, the general ideas and principles gained from the work up to about 1934, giving what may be called the "Background" to the more recent developments; the second, the theoretical and experimental work since that date connected with the first major theme, "Intrinsic Magnetization"; and the third, the more recent work connected with the second major theme, which can be indicated by the title "Magnetization Curves". The field, however, proved so extensive that it could not be covered in the manner which seemed appropriate in a single report of reasonable length. The present report, therefore, covers only the first two of the three parts envisaged. The "Background" is sketched in §2, and the more recent work on "Intrinsic Magnetization" is surveyed in §§3 and 4. It is hoped that the third part may be presented in a further report dealing with recent work on magnetization curves and associated phenomena in the next volume of the Annual Reports.

For the most part the survey moves in a middle air where theory and experiment come into close contact with each other. Although, in general, no details can be given either of experimental methods and procedure or of theoretical arguments, an indication is given, where appropriate, of the premises of theoretical treatments, since otherwise the statement of theoretical conclusions is of little value. It should be stated, however, that it is not always easy to discover what are the premises of theoretical treatments; nor, occasionally, what, if any, are the conclusions.

§ 2. BACKGROUND

2 (i). *Introductory*

This sketch of the background to the more recent developments in ferromagnetism is principally concerned with work up to about 1934, and specific references to papers are given only where later work is mentioned. Fuller surveys of most of the topics touched upon, as well as detailed references, are given in one

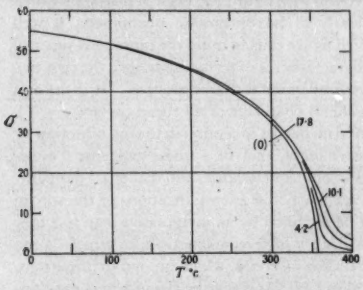


Figure 2.1. Nickel. Variation of magnetization with temperature. σ, specific magnetization. The numbers on the curves give the applied field in kilo-oersteds. The curve labelled (0) is the estimated curve, obtained by extrapolation from the high field values or otherwise, for the spontaneous magnetization.
(*After Weiss and Forrer*, 1926.)

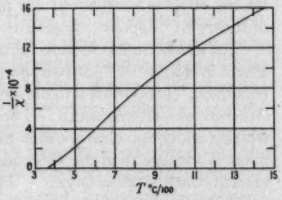


Figure 2.2. Nickel. Variation of susceptibility above the Curie temperature.
(*After Fallot*, 1944.)

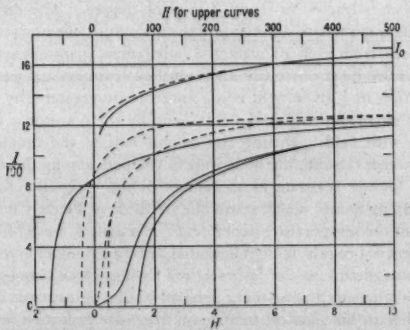


Figure 2.3. Initial and cyclic magnetization curves of mild steel (C 0·04, Mn 0·40 %). Full curves, as cast. Broken curves annealed from 900° c. (Data from *Int. Crit. Tables*.)

or other of the books listed in the bibliography, but since ferromagnetism does not figure prominently in most of the standard general text-books, and the reading of more specialized works often remains a good intention, it is thought that the general survey given here may not be without value to a number of readers.

It may be well to draw attention at the outset to the central experimental subject matter of ferromagnetism by three figures showing the variation with temperature of the saturation magnetization of a typical ferromagnetic below the Curie point (figure 2.1), the variation of susceptibility above the Curie point (figure 2.2) and a representative curve for the variation of magnetization with field (figure 2.3). Remembering the variety in detail in the behaviour of different ferromagnetic materials, and, in connection with figure 2.3, even of materials which are chemically identical, the complexity of ferromagnetic phenomena is well exemplified by these figures. The third figure calls to mind the more obvious, and the technologically more important characteristics of ferromagnetics; the first two the less obvious but theoretically more fundamental characteristics. It is the aim of theory to "explain" at least the general character of all these curves.

The modern outlook on ferromagnetism had its inception in the introduction by Pierre Weiss, in 1907, of two speculative ideas, that of a quasi-magnetic "molecular field", proportional to the intensity of magnetization, and that of a "domain" structure of ferromagnetics. There are few better exemplifications of the way in which a wider and clearer vision may be obtained by an imaginative leap to a new point of vantage than the subsequent history of ferromagnetism, to which Weiss himself contributed so largely. These ideas of Weiss, which proved so important, can perhaps be best introduced by recalling, very summarily, the previous development of the subject.

2(ii). *The nineteenth century*

The puzzle of ferromagnetism is a very old one—it is well over two thousand years ago that the attraction of iron by "magnets" was first observed and marvelled at; it is barely a hundred years ago that it was first convincingly shown by Faraday that all bodies are attracted or repelled in some degree. The difference in time reflects the great quantitative difference of the magnetic behaviour of iron and a few other substances from that of attracted substances, paramagnetics, generally. Ferromagnetics are most obviously distinguished from normal paramagnetics by the magnetization in a given field being enormously greater—by a factor of the order of 10^6 in small and moderate fields—and by the non-linear variation of the magnetization with field. During the second half of the nineteenth century, particularly towards the end, methods were developed, among them many of those still standard, for the accurate measurement of the magnetic characteristics of materials. Among names which stand out are those of Pierre Curie, whose great work of 1895 on the temperature dependence of magnetic susceptibility of a wide range of substances brought to light essential distinguishing characteristic of dia-, para- and ferromagnetics; and of James Alfred Ewing, whose many-sided investigations on ferromagnetism, among them detailed studies of magnetic hysteresis, were brought together in his classical treatise on *Magnetic Induction in Iron and other Metals* (1892, 1900). Some progress had been made before 1900 towards a theoretical interpretation, from an atomic standpoint, of magnetic properties, in that diamagnetism was attributed to the induction by the field of molecular currents (which at once explained the repulsive effects), and paramagnetism to the partial alignment in the direction of the field of atoms or molecules possessing a permanent magnetic moment due to circulating currents. Explanations of

ferromagnetism were also attempted in terms of the behaviour of aggregates of atomic magnets, but the relation between ferro- and paramagnetism remained very vague; indeed, the partial apparent success of a treatment of hysteresis on this basis, without reference to paramagnetism, obscured fundamental issues and held up for many years a proper appreciation of the essential character of ferromagnetism.

2(iii). *Langevin's theory of paramagnetism*

During the last decade of the nineteenth century a firm experimental basis was provided for the electron theory of matter, so extensively developed by Lorentz, and a new stage in magnetism is reached with the application of the theory to dia- and paramagnetism by Langevin in 1905. Langevin's treatment differs in two respects from treatments given earlier: first, it was based on a more precise, but not too particularized, picture of the electronic structure of the atom; and second, the problem was treated statistically in a clear and straightforward manner. (That some of the assumptions were, strictly, unjustifiable, is a criticism of the adequacy of classical assumptions.) The case considered is that of a substance in which the atoms (or molecules) are sufficiently widely separated for the mutual magnetic effects to be negligible, as in a gas. A magnetic moment is associated with the orbital motion of each electron in the atom. The resultant magnetic moment of the atom may be zero, when the effect of a field is simply to modify the motion of the electrons by the superimposition of a precession (as indicated by Larmor's theorem), giving an induced magnetic moment in the opposite direction to the field. The orders of magnitude, and the temperature independence of the susceptibilities of "normal" diamagnetics are satisfactorily accounted for. If the atoms have a resultant moment there is in addition to the diamagnetic effect a usually much larger paramagnetic effect. The atoms may all be supposed to have the same resultant moment μ. The energy of the atom depends on the orientation of the magnetic moment in the field; for an angle θ between μ and H the magnetic energy is $-\mu H\cos\theta$. By the usual Maxwell-Boltzmann statistics, in a state of equilibrium, however, this may be attained, the number of atoms with axes in a direction θ, per unit solid angle, will be proportional to $\exp(+\mu H/\boldsymbol{k}T)$. A straightforward derivation gives the well known Langevin expression for the mean resolved moment, $\bar{\mu}$, in the field direction,

$$\bar{\mu}/\mu = \coth(\mu H/\boldsymbol{k}T) - (\boldsymbol{k}T/\mu H). \qquad \ldots\ldots(2.1)$$

For large values of $\mu H/\boldsymbol{k}T$ this approaches unity (i.e. the magnetic moments tend to parallelism with the field); for small values of $\mu H/\boldsymbol{k}T$ (which in practice, at room temperature, usually does not exceed 0·01 for a field of 10 000 oersted),

$$\bar{\mu}/\mu = \mu H/3\boldsymbol{k}T, \qquad \ldots\ldots(2.2)$$

or, for the mass susceptibility,

$$\chi = n_{\mathrm{m}}\bar{\mu}/H = n_{\mathrm{m}}\mu^2/3\boldsymbol{k}T, \qquad \ldots\ldots(2.3)$$

where n_{m} is the number of atoms per unit mass.

Langevin thus obtained a theoretical derivation of Curie's law of the inverse dependence of paramagnetic susceptibility on temperature, followed closely by oxygen, and at least roughly by many paramagnetics over considerable temperature ranges. The Curie law may be written

$$\chi = C_{\mathrm{m}}/T, \qquad \ldots\ldots(2.4)$$

and when this law is obeyed, the magnetic moment of the "carriers"—a convenient term—may be estimated, using (2.3), from the Curie constant, C_m. The theory indicates that although ordinarily the magnetization of a paramagnetic varies linearly with the field, for sufficiently high values of H (or, more precisely, of H/T), saturation would be approached, as has been confirmed by low temperature measurements on paramagnetic salts. The estimated saturation magnetization, $n\mu$, may, however, be hundreds of times greater than the highest magnetization, $n\bar{\mu}$, ordinarily observed. Of fundamental importance in connection with ferromagnetism is the fact that the saturation magnetization estimated in this way for typical paramagnetics (usually salts containing transition element ions) is of the same order as the quasi-saturation magnetization observed at room temperature or lower for typical ferromagnetics, such as iron, cobalt or nickel. In other words, ferromagnetics do not differ from paramagnetics in the order of magnitude of the magnetic moments of the elementary carriers. The distinguishing characteristic of ferromagnetics is that a close approach to parallel alignment of the magnetic axes of the carriers is attained for small values of H/T (e.g. for a field of a few hundred oersted, or even much less, at room temperature), suggesting that in ferromagnetics there may be an effective field very much greater than the applied field. The Weiss treatment may be taken up from this point.

2 (iv). *The Weiss theory of ferromagnetism*

(a) *The molecular field hypothesis*

Weiss assumed that the magnetic effects of the interactions between atoms could be represented by a quasi-magnetic molecular field proportional to the intensity of magnetization; denoting this Weiss field by H_w,

$$H_w = N_w I, \qquad \ldots\ldots (2.5)$$

where N_w is the molecular field constant. Formally the equation is identical with that for purely magnetic interaction, for which, however, the constant would be of the order $\frac{4}{3}\pi$, while the values required for N_w, to fit the experimental data, may be hundreds of times greater, and may be negative. The origin of molecular fields of the magnitude required was in fact unknown (and remained unknown until 1928), and the justification for the assumption lies in the extent to which it leads to a better coordination of the phenomena. With an external field H_e, the total field is

$$H = H_e + H_w = H_e + N_w I, \qquad \ldots\ldots (2.6)$$

and the molecular field theory, in its original form, is simply the development of the consequences of (2.1) and (2.6). The main consequences are well known. A substance to which the equations apply, with n_m carriers of moment μ per unit mass, will have a mass susceptibility above a temperature θ given by

$$\chi = C_m/(T-\theta), \qquad \ldots\ldots (2.7)$$

where $C_m = n_m\mu^2/3k$, $\theta = N_w\rho C_m$.

Below θ the substance will be spontaneously magnetized without any application of an external field to a degree increasing rapidly at first and then more slowly to a complete saturation value at absolute zero. A "universal" curve similar to those given later may be obtained by plotting the reduced magnetization, $\bar{\mu}_T/\mu_0$, against the reduced temperature T/θ. An applied field will increase the magnetization only very slightly, except near the Curie point.

As a first approximation the Curie-Weiss law (2.7) covers the behaviour of ferromagnetics above the Curie temperature tolerably well. It may be mentioned in passing that many paramagnetic salts follow this law remarkably closely over considerable ranges of temperature, but for these salts θ is best regarded simply as a parameter in the empirical equation representing the results, for there are usually very wide deviations from the law as the temperature θ is approached. Negative as well as positive values of θ are found, with numerical values usually less than 90° (to be compared with a Curie temperature for iron, for example, of about 1040° K.). In so far as the same equation (2.7) holds both for ferromagnetic metals above the Curie point, and for paramagnetic salts, it can be said that, formally, the Weiss molecular field treatment applies to both, but it does not follow that the physical origin of the field is the same in the two cases; with minor reservations it has turned out, in fact, to be completely different.

(b) *The domain hypothesis*

At first sight the implications of the molecular field treatment as to spontaneous magnetization of ferromagnetics below the Curie point would not appear very promising. It is true that ferromagnetics become strongly magnetized in moderate fields, and permanent magnets show that a state of spontaneous magnetization can occur; but most ferromagnetics can exist quite happily in an apparently unmagnetized state in the absence of an applied field. The suggestion of Weiss is that an apparently unmagnetized specimen is in fact spontaneously magnetized to the degree appropriate to the temperature as given by the molecular field treatment, but with the direction of magnetization differing from region to region in the specimen. These regions of uni-directional magnetization, commonly called domains, may be small enough to be compatible with uniformity of magnetization (or non-magnetization) on the ordinary scale, and still contain a large number of atoms (for iron, for example, a cube of side 0·001 mm. contains about $8{\cdot}5 \times 10^{10}$ atoms). It was at first thought likely that the domains might be identified with micro-crystals, and a rough treatment of the character of the magnetization curves to be expected was given by attributing to the domains properties similar to those deduced for "elementary" crystals of pyrrhotite (which had previously been studied by Weiss). These crystals have a single "easy" axis of magnetization, along which they are spontaneously magnetized to the saturation value. The sense of magnetization changes at a field of about 15 oersted along this axis (giving rectangular hysteresis loops); along directions at right angles saturation is approached only in fields of several thousand oersted. By considering an aggregate of domains with similar characteristics, and with easy axes orientated at random, the general form of ordinary hysteresis curves is roughly reproduced, but it is to be noted that this treatment in no sense "explains" hysteresis. The treatment of domains as pyrrhotite-like should be regarded as no more than illustrative of possibilities. It was not until 1918 that methods were developed which made possible investigations of the magnetic properties of single crystals of the ferromagnetic metals. It then became clear that domains were not to be identified with single crystals or crystal grains, but that the concept of domains as bounded, usually "small", regions over which the spontaneous magnetization is uni-directional, was as necessary in accounting for the behaviour of single crystal as of polycrystalline material.

Effectively direct evidence of a domain structure of ferromagnetics is provided by investigations of the discontinuous character of the magnetization process, first revealed by Barkhausen's experiments of 1919. Before then, however, there was very strong indirect supporting evidence for the general Weiss viewpoint. According to this, the effect of an applied field, unless it is very large, is to change the direction of the domain magnetization without appreciably changing its magnitude, so that the "intrinsic" magnetization, approximately equal to the ordinarily measured saturation magnetization at the temperature, is practically independent of the apparent magnetization. The energy associated with the intrinsic magnetization, $-\frac{1}{2}N_w\rho I_T^2$, where I_T is the intrinsic magnetization at temperature T, increases with temperature as I_T decreases, and will give rise to an "additional" specific heat which attains a sharp maximum at the Curie point. On this basis Weiss was able to account, at least qualitatively, for the observed specific heat "anomalies" of a number of ferromagnetics. Further evidence of a similar kind both for the molecular field and the domain hypothesis was provided from measurements on the high field magneto-caloric effect. This effect, a reversible change in temperature accompanying change in magnetization, was discovered by Weiss and Piccard in 1917 in measurements on ferromagnetics in high fields over a range of temperatures; apart from its special characteristics in ferromagnetics, the effect is the same as that applied in the production of very low temperatures by the adiabatic demagnetization of paramagnetics. The essential characteristic of the effect, in the present connection, is that the temperature changes only very slightly as the apparent magnetization changes from zero up to nearly the ordinary saturation value, but it then changes rapidly, ultimately approximately linearly with the square of the magnetization. The relatively large change—changes of the order of 1° have been observed for nickel near the Curie point—can be associated with an increase in the "true" or "intrinsic" magnetization in sufficiently high fields. The zero or very small change in lower fields gives strong support for the view that here the change in apparent magnetization is due almost entirely to changes in direction, but not in magnitude, of the intrinsic magnetization of the domains. Extrapolation of the upper linear parts of the high field ΔT, I^2 curves to zero ΔT gives perhaps the most precise means of determining the zero field intrinsic magnetization, that is, the spontaneous magnetization, at the higher temperatures.

The two hypotheses of Weiss, which, with their development, constitute his theory of ferromagnetism, have been dealt with at what may seem undue length in an out-of-date setting, partly because they underlie so much of the later work, and partly because the magnitude of Weiss's contribution in bringing order out of chaos is seldom fully appreciated. But the shortcomings of the theory should not be overlooked. Some of these, notably the purely formal rather than physically grounded nature of the basic assumption in the molecular field treatment, are obvious, and could be removed only as a result of later developments of more general theory. Apart from this, however, the coordination effected is much more imperfect quantitatively than is generally supposed. The experimental form of the "saturation" magnetization temperature curve, for example, differs very considerably from that given by the theory, and, again, alternative methods of deriving the molecular field constant, N_w, and the carrier moment, μ, from the

experimental results often yield widely different values. Difficulties of this kind are partly, but by no means entirely, removed by a straightforward re-expression of the Weiss theory on a quantum rather than on a classical basis.

2(v). *The quantum theory and elementary carriers*

The magnetic moments per atom of iron and nickel, as deduced from the low temperature saturation magnetizations extrapolated to absolute zero, are very closely in the ratio of 11 to 3, and Weiss suggested, in 1911, that elementary magnetic moments were all integral multiples of the highest common sub-multiple of these two moments, other evidence having made a suggestion of this kind at least plausible. The Weiss "magneton" ($\mu_w = 1{\cdot}87 \times 10^{-21}$ erg. gauss^{-1}, corresponding to a gram molecular value, M_w, of $1{\cdot}125 \times 10^{3}$ erg. gauss^{-1} mole^{-1}) has been widely adopted up till very recently as a unit in the expression of the values of elementary magnetic moments in both paramagnetics (usually as deduced from the Langevin-Weiss treatment outlined above) and ferromagnetics, and is mentioned here, for this reason. In many respects the suggestion was extremely valuable, particularly in stimulating accurate magnetic measurements, and making possible a convenient numerical expression of experimental results, but even as a purely empirical unit the Weiss magneton has now long outlived its usefulness.

The quantum theory, as developed by Bohr in his famous paper of 1913, immediately suggests a "natural" unit of magnetic moment. In so far as the angular momentum of an electron in orbital motion in an atom is an integral multiple of $\boldsymbol{h}/2\pi$, the associated magnetic moment would be expected, on a quite general electron theory basis, to be an integral multiple of $e\boldsymbol{h}/4\pi\boldsymbol{m}c$. The value of this unit, the Bohr magneton, is given by $\mu_B = 9{\cdot}27_3 \times 10^{-21}$, or, for the gram-molecular unit, $M_B = 5{\cdot}58_5 \times 10^{3}$. (It may be noted that $\mu_B/\mu_w = 4{\cdot}97$. That the ratio is nearly integral is a bare numerical fact without other significance.) Later developments of the quantum theory suggested that not only was the angular momentum quantized, but also that the resolved value of the angular momentum, and hence of the associated magnetic moment, could assume only certain discrete values in the direction of a magnetic field (Debye, Sommerfeld, 1916). It must suffice to say here that this at first startling suggestion received abundant experimental confirmation, both direct (in experiments of the Stern-Gerlach type) and indirect (particularly from the study of the Zeeman effect), and that it now appears simply as a particular conclusion from much more general theoretical premises. A drastic revision of the classical formulae for paramagnetism (and ferromagnetism) is entailed, for instead of finding the distribution, and averaging, over a continuous set of possible orientations as in deriving the Langevin expression (2.1), it is necessary to consider a discrete set. A rather detailed knowledge of the quantum specification of atomic and molecular states is an indispensable preliminary to an appreciation of the modern theoretical treatment of paramagnetism (more particularly of salts), but, for a special reason, it is largely irrelevant to the consideration of ferromagnetism.

Towards the end of the first phase of the quantum theory (1925) a wide coordination of facts relating to atomic structure and spectroscopic characteristics was attained by attributing to the electron an intrinsic spin, with a "mechanical" moment equal to $\frac{1}{2}$, in Bohr units (i.e. a moment of $\frac{1}{2}(\boldsymbol{h}/2\pi)$), and an associated

magnetic moment of 1 Bohr unit (i.e. $eh/4\pi mc$), the ratio of the magnetic to the mechanical moment being double that for the orbital motion of the electron.* The ratio of magnetic to mechanical moment, when both the moments are expressed in Bohr units, is denoted by g (the Landé "splitting" factor which enters into the expressions for the splitting of lines in the Zeeman effect). For electron spin, $g=2$, for purely orbital motion, $g=1$; for the resultant moment of an atom to which both spin and orbital moment may contribute, other (calculable) g values hold, not necessarily lying between 1 and 2. Now the ratio of the mechanical to the magnetic moment of the carriers in ferromagnetics may be determined almost directly by measurements of the effective field, shown by change of magnetization, associated with rotation or of the rotational impulse associated with change of magnetization, effects first definitely established experimentally by Barnett (1914) and by Einstein and de Haas (1915) respectively. The experiments are difficult and delicate, the quantities actually to be measured being very small, and the records of subsequent investigations provide some remarkable examples of skilful experimental technique. The details of this work must be passed over here, but many surveys have been given, including two recent ones (1940, 1944) by Barnett, whose outstanding work in this field has extended over more than thirty years. There is general agreement among all investigators that the g value for the carriers of the magnetic moment in ferromagnetic metals and alloys is very close to 2. Barnett, however, finds consistent differences from 2, outside the range of his experimental uncertainties, a mean value for all the materials studied being 1·90, and particular values for iron and nickel being 1·94 and 1·90, with an estimated error of within 1 %. The conclusion to be drawn from all this work is that change in magnetization of a ferromagnetic metal or alloy is almost entirely due to change of orientation of electron spins. At least to a close approximation, ferromagnetics may be treated as though electron spins only were involved. The orbital moment, however, is not completely "quenched" and that it may play a part, even an important part, in some ferromagnetic effects is not ruled out. Orbital moment may be of much greater importance in "non-metallic" ferromagnetics, as is shown conclusively by the g value of approximately 0·63 found by Coeterier (1935) for pyrrhotite. There is no *a priori* reason for supposing that in metallic ferromagnetics the almost complete predominance of the spin effect is peculiar to temperatures well below the Curie point; direct evidence that it is not is provided by gyromagnetic measurements on a nickel-copper alloy (56% Ni) above its Curie point by Sucksmith (1935) who obtained an approximate g value of 1·9.

The importance of work on the gyromagnetic effects in showing that the elementary magnets in ferromagnetics are primarily electron spins can hardly be over-emphasized. In one respect it introduces a considerable simplification in the approximate theoretical treatment of ferromagnetism, as will be clear from the modified molecular field treatment, presented below. At the same time it

* For an electron in a state specified by an angular momentum quantum number j, the quantum mechanical treatment allows of a "sharp" value, equal to $j(j+1)(h/2\pi)^2$, for the square of the angular momentum and also of a sharp value for the resolved component in a particular direction, such as might be determined physically by a magnetic field, but not of a sharp value at right angles to this direction. The maximum resolved angular momentum in any direction is $j(h/2\pi)$, as in the earlier treatment, and to provide a link with this, the angular momentum, in units of $h/2\pi$, is often stated loosely, as here, to be j, though strictly the value should be given as $\sqrt{\{j(j+1)\}}$. For electron spin, $j=s=\frac{1}{2}$.

introduces, or makes apparent, special difficulties. Thus the effective number of electron spins per atom, as determined from the saturation magnetization extrapolated to absolute zero, is not integral (the number is approximately 2·21 for Fe, 0·60 for Ni). This gives a clear-cut exemplification of the kind of problem with which the theory of ferromagnetism is confronted.

2(vi). *Interchange interaction*

With the advent of quantum mechanics not only was a self-consistent theoretical scheme developed which incorporated all that was of permanent value in the earlier quantum theory, but also many long-standing problems were opened to theoretical attack which had been completely outside the scope of any earlier theoretical treatment. Among these is the problem of molecule formation, or perhaps more generally of chemical valency or binding forces, empirically co-ordinated, perhaps, but hitherto as mysterious qualitatively and quantitatively as the Weiss molecular field. The fundamental explanation of the origin of this field was given in 1928 by Heisenberg, and also at about the same time independently by Frenkel, in terms of interchange interaction effects essentially similar to those brought to light a year earlier in the quantum-mechanical treatment of the hydrogen molecule by Heitler and London. As the Heitler-London method is a standard theme in books on quantum mechanics, it is unnecessary here to do more than indicate, as briefly as possible, those points relevant to a general understanding of ferromagnetism.

Let $u_a(1)$, $u_b(2)$ be the wave functions for two separated hydrogen atoms, each in its lowest orbital state (leaving the electron spins for the moment out of consideration), 1 and 2 referring to the electrons associated, in the separate atoms, with the nuclei a and b respectively. A straightforward application of the Schrödinger equation shows that there are two states for the composite system with energies differing by an amount which depends on the internuclear distance. Neglecting non-orthogonality "corrections" which are large for the hydrogen molecule but, as shown by Van Vleck (1936), probably relatively unimportant in the applications to ferromagnetism, the energy values for the system may be written

$$E = 2\epsilon = 2\epsilon_0' \pm J_0. \qquad \ldots\ldots (2.8)$$

Here ϵ_0' includes the unperturbed energy of the separated atom, and also the energy associated with the Coulombian attractions and repulsions, and

$$J_0 = \int\int u_a(1)\, u_b(2)\, V_{ab}\, u_b(1)\, u_a(2)\, dv_1\, dv_2, \qquad \ldots\ldots (2.9)$$

where

$$V_{ab} = \tfrac{1}{2}e^2\left(\frac{2}{r_{ab}} + \frac{2}{r_{12}} - \frac{1}{r_{a1}} - \frac{1}{r_{b2}} - \frac{1}{r_{a2}} - \frac{1}{r_{b1}}\right),$$

in an obvious symbolism. The energy function corresponding to the positive sign in (2.8) has a minimum at a value of r_{ab} to be identified with the equilibrium internuclear distance in the hydrogen molecule. The calculations, which, it need hardly be said, are very troublesome even for hydrogen, show that J_0 is negative over a wide range of r_{ab} values, and in particular for the equilibrium separation, where its numerical value is large, accounting for most of the energy of formation of the H_2 molecule (about 4·7 electron volts). In this state the wave function is symmetrical (of the form $u_a(1)\, u_b(2) + u_a(2)\, u_b(1)$), in the higher energy state

antisymmetrical. Now the Pauli exclusion principle can be formulated in the requirement that the complete wave function must be antisymmetrical in the electron coordinates, including the spin coordinates; a symmetrical orbital function must therefore be associated with an antisymmetrical spin function. There is just one state of this type, corresponding physically to an anti-parallel alignment of the spins, the resultant spin and associated magnetic moment being zero. In the higher energy state of the hydrogen molecule, antisymmetry requires a symmetrical spin function; the spins are aligned parallel to each other, and there are in all three states corresponding to the three possible resolved values, 1, 0 and -1, of the resultant spin 1.

The effect is as though the electron spin magnets had a mutual potential energy proportional to the (quantum-controlled) scalar product of their magnetic moment vectors, but the proportionality factor involves J_0 and may be enormously greater than for any purely magnetic interaction. The effect has, in fact, nothing to do with magnetic interaction at all, being rather of electrostatic origin, and strictly it is not even quasi-magnetic, for the potential, unlike that of magnetic dipoles, is isotropic; the spin arrangement is, in a sense, simply an indicator of the symmetry character of the orbital functions. With the unperturbed wave functions appropriate to the treatment of the interaction of normal hydrogen atoms (1s functions), the interchange interaction integral is, as has been seen, negative. If, however, J_0 were positive, the two energy states would change their rôles, and the lower energy state would be associated with parallel alignment of spins. This is the essence of Heisenberg's proposed interpretation of the Weiss molecular field. The development of the theory involves on the one hand a consideration of the conditions under which J_0 may be positive and on the other an appropriate statistical treatment of an aggregate of interacting atoms.

It is important to bear in mind that the perturbation method used is based on the approximation of treating the interaction between the atoms as "weak". The results obtained are lacking in quantitative accuracy for the hydrogen molecule, and the inaccuracy is likely to be much more pronounced in the treatment of the ferromagnetism of solids. The method does, however, bring out clearly significant features of the physical situation.

Except in the simplest cases there is little hope of a quantitative theoretical calculation of the interchange interaction of sufficient accuracy to be useful, but following the qualitative considerations due to Heisenberg, Bethe and others, an indication can be given of the conditions favourable to a positive rather than a negative interchange integral. Referring to (2.9), if the wave functions u_a and u_b have no nodes in the region where the overlap is appreciable, the product of the wave functions may be assumed as an approximation to be everywhere positive. This holds for 1s wave functions (as for hydrogen), and may also hold effectively in certain circumstances for p, d, f ... wave functions, particularly for those of the type 2p, 3d, 4f.... The integral J_0 will then be positive if the contributions from the repulsive terms involving r_{ab} and r_{12} exceed those from the remaining attractive terms. This is favoured by the wave functions being small in the neighbourhood of the nuclei, and large in those regions for which r_{12} is small, that is in the neighbourhood of the mid-plane. A function such as $u_a(1)$, for example, should be (1) small near a, (2) large mid-way between a and b, and (3) small near b. Functions

other than s functions satisfy the first condition, and, other things being equal, the maximum moves outwards with increasing l. In view of conditions (2) and (3) it is tempting to say (knowing the facts) that these theoretical considerations suggest that d and f functions will be more favourable than p functions to ferromagnetism, with f possibly less favourable than d. All that can justifiably be said is that these considerations provide some theoretical basis for a useful guiding rule, due to Slater, to the effect that with incomplete electron groups (d or f) in atoms the magnitude of the interchange interaction is dependent on the ratio of the interatomic distance to the effective radius of this group, and that, over most of the range involved, it increases algebraically as this ratio increases. Eventually, with large interatomic distances, it must decrease numerically to zero. The relevant experimental facts will be considered later.

On the statistical side, Heisenberg's very elaborate treatment showed that a positive interchange interaction could have effects similar to those of a Weiss molecular field, among these being spontaneous magnetization below a critical temperature. In short, the essential characteristic of a ferromagnetic can be explained in terms of positive interchange interaction. In detail, however, the results of the statistical treatment were almost completely at variance with the facts. It appeared, for example, to the approximation to which the work was carried, that ferromagnetism would not occur if the number of nearest neighbours, z, round each atom were less than 8; that with $z=8$ spontaneous magnetization would appear only in the immediate neighbourhood of the Curie point; and that even with $z=12$ it would disappear at a lower temperature. These peculiar results can be traced partly to an assumption that interchange interaction effects are negligible except between nearest neighbours, and partly to a particular form assumed, with unavoidable arbitrariness, for the energy distribution of states with a given total spin. While Heisenberg's treatment may now properly be considered as belonging to the "background", the subsequent improvements, and also alternative methods of approach to the general problem, are best taken up in connection with the later developments in §3. It will, however, be appropriate to give here, in connection with a modified molecular field treatment, a simple indication of how positive interchange interaction can lead to ferromagnetism.

2(vii). *Modified molecular field treatment*

The molecular field treatment may readily be modified to allow for the discrete orientation of the carriers, effectively electron spins, in place of the continuous orientation of the original Langevin-Weiss theory. An electron spin will set itself parallel or anti-parallel to the field given by (2.6), the energies being

$$\epsilon = \pm\mu H = \pm\mu(N_w I + H_e). \qquad \ldots\ldots(2.10)$$

There follows, in place of (2.1) the simpler result,

$$\bar{\mu}/\mu = \tanh(\mu H/kT). \qquad \ldots\ldots(2.11)$$

This gives spontaneous magnetization below a temperature, θ, and above, a mass susceptibility given by

$$\chi = C_m/(T-\theta), \qquad \ldots\ldots(2.12)$$

where

$$C_m = n_m\mu^2/k, \qquad \theta = N_w\rho C_m, \qquad C_m = N_w n_v(\mu^2/k),$$

where μ is the Bohr magneton, and n_v the number of effective independent spins

per unit volume. It is convenient for the general expression of the results to use non-dimensional magnitudes as in the two "universal" curves, shown in figure 2.4, in which the abscissa is the reduced temperature, T/θ, and the ordinate, for $T \leqslant \theta$,

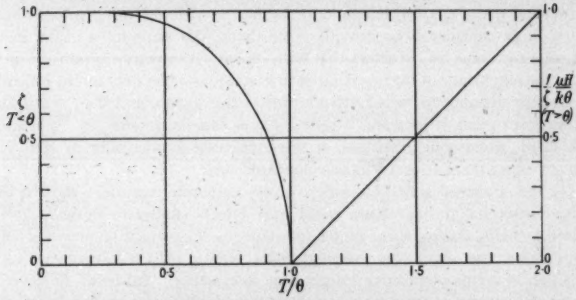


Figure 2.4. Intrinsic magnetization below, and paramagnetism above the Curie point. Molecular field treatment for electron spins.
ζ, relative magnetization.
μ, Bohr magneton.

the relative magnetization, ζ, equal to the ratio of the excess parallel spins in one direction to the total number, and equivalent to $\overline{\mu}/\mu$, and for $T \geqslant \theta$, a suitably reduced value of the inverse of the susceptibility.

In so far as the curves of figure 2.4 give a satisfactory representation of the magnetic characteristics of ferromagnetics, a ferromagnetic would be specified by only two "parameters", for which the choice would ordinarily be the Curie temperature, θ, and, as may be most convenient, the total number of effective electron spins per unit volume, per unit mass or per atom. (The treatment as given applies to independent spins with $j = \frac{1}{2}$. It can readily be modified for coupled spins with $j = 1, \frac{3}{2}$ etc.) The general fit with the experimental results for the ferromagnetics which have been most thoroughly investigated is very much better (the least that could be expected) than with the purely classical treatment. The spontaneous magnetization curve for independent electron spins as carriers (carriers with $j = \frac{1}{2}$) agrees better with experiment than do any other simple quantum curves ($j = 1, \frac{3}{2}$ etc.) all of which lie lower, between the $j = \frac{1}{2}$ and the classical curve. The agreement, however, is by no means perfect. For small values of T/θ the experimental points are usually slightly below, and for values approaching 1, more markedly above the curve. In many other ways, too, the agreement in detail is unsatisfactory. The number of effective spins, as (accurately) estimated from the saturation magnetization usually differs considerably from the number as (roughly) estimated from the temperature dependence of the paramagnetism (for nickel, for example, the estimated numbers per atom are 0·606 and about 0·86 respectively). Further, the relation between $1/\chi$ and T above θ is by no means strictly linear; at least near θ, a curvature, concave upwards, is well established. Finally, it is not to be supposed that consistent results are in general obtained in deducing the value of the molecular field constant, N_w, from θ on the one hand and from the various magneto-thermal effects on the other.

The "theoretical" curve for the additional specific heat consistent with the intrinsic magnetization curve of figure 2.4 is shown in figure 2.5. The "jump"

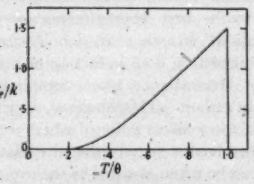


Figure 2.5. Additional specific heat associated with decrease in intrinsic magnetization. Molecular field treatment for electron spins. c_e, additional heat capacity per effective electron.

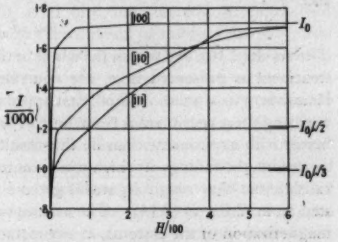


Figure 2.6. Magnetization curves for single crystals of iron. (*After Honda and Kaya*, 1926.)

in the neighbourhood of the Curie point as estimated from experiment is usually at least roughly of the order indicated by this curve, but it is much less abrupt; for nickel there is a "tail" extending over 30° or more, and apart from this, there is a fairly constant excess specific heat above the Curie point (of the order of 1 cal. deg^{-1} mole^{-1}) completely unaccounted for by the present treatment. None of these points is trivial, in the sense that it calls merely for a slight and obvious emendation of the theoretical treatment; at the same time none of them invalidates the use of the semi-theoretical results, or the curves of figures 2.4 and 2.5 as probably the most nearly satisfactory simple standards of reference for the consideration of experimental results.

It remains to show how the molecular field and the Curie temperature may be interpreted in terms of interchange interaction. The essentials are perhaps most readily conveyed by a simple, if somewhat crude, approach. Suppose that the atoms of the ferromagnetic, each with one effective electron spin, interact appreciably with z surrounding atoms, and that the interchange interaction integral for each of the z pairs of atoms, including the central atom, is J_0. Of the interaction energy involved, half may be associated with the central atom. If x and y of the z atoms have spins pointing to the right and to the left respectively, then, according to whether the spin of the central atom is pointing to the left or the right, the energy associated with it will be

$$\epsilon = \pm \tfrac{1}{2}(x-y)J_0. \qquad \ldots\ldots(2.13)$$

Now assuming that the magnetization of the z atoms is a fair sample of the magnetization as a whole,

$$(x-y)/z = I/I_0,$$

and

$$\epsilon = \pm \tfrac{1}{2}(zJ_0/I_0)I. \qquad \ldots\ldots(2.14)$$

Comparison with (2.10) and (2.12) gives at once

$$N_w = \tfrac{1}{2}(zJ_0/\mu I_0) = \tfrac{1}{2}(zJ_0/n_v\mu^2), \qquad \ldots\ldots(2.15)$$

and

$$\theta = zJ_0/2k. \qquad \ldots\ldots(2.16)$$

The result (2.16), and with it the whole of the results of the modified molecular field treatment as presented here, are equivalent to the first approximation of the Heisenberg treatment. This "first approximation", in spite of its deficiencies in covering the experimental behaviour of ferromagnetics, does in fact agree much better with experiment than do the results of Heisenberg's more complicated treatment, the failings of which were pointed out earlier. This, however, does not validate the very rough argument given above, the weakest point of which is the step from (2.13) to (2.14). The assumption involved is stated—that the relative magnetization of the z atoms, at any instant, can be taken as equal to the average relative magnetization of the material in bulk—and it will be more nearly valid the greater the value of z. The smaller the value of z, the more important are the effects of fluctuations, and the first approximation results may be profoundly modified. Some modification is required, but the modification introduced by supposing that effective interchange interaction is restricted to that between nearest neighbours, even if it is of the right kind, is far too drastic.

2 (viii). *Single crystals*

The complicated behaviour of ferromagnetics in low and moderate fields cannot be accounted for in terms of interchange interaction effects alone. An obvious approach to the problems presented by the behaviour of ordinary polycrystalline metallic ferromagnetics is by the examination of the magnetic properties of single crystals, but at the time when Weiss's work on pyrrhotite and magnetite indicated the desirability of such investigations (see §2 (iv) (b)), methods for growing metallic crystals of sufficient size had not been developed. It has now been found that large single crystals of metals may be produced by controlled cooling of the molten metal, or by prolonged heating at a suitable temperature followed by slow cooling of a polycrystalline rod which has been stretched. The first method is applicable to nickel, but, owing to the various phase changes, not to iron. The pioneer magnetic studies of single metallic crystals were made by Beck (1918) and Webster (1925) and since then extensive investigations have been carried out, particularly by the Japanese, on ferromagnetic alloys as well as on the metals, in some cases over a wide range of temperatures. Torque methods have been employed, using suitably cut discs, and also ballistic methods. The general character of the results is well illustrated by the frequently reproduced curves obtained by Honda and Kaya (1926) for iron, with the field along the cube edge, face diagonal and cube diagonal directions (figure 2.6). The details of the behaviour in low fields, difficult to determine with accuracy, and markedly dependent on the perfection of the specimens, are not of great importance. On the scale of the diagram, it may be said that along the [100] direction the magnetization rises steeply almost to the saturation value ($I_0 = 1720$); along the [110] and [111] directions the initial rise is equally steep, but comparatively sharp bends occur at I values of about 1200 and 1000 respectively. Approximate

saturation is reached in the three directions at about 70, 650 and 450 oersted. The cube edge directions are conveniently described as "directions of easy magnetization". In the medium field range the field and magnetization are not in general co-directional except in the axial directions already specified, and intriguing periodic curves are obtained for the variation of the parallel and transverse components of the magnetization with the angular direction of the field in various crystal planes.

Little progress can be made in the interpretation of the single crystal results without first assuming a domain structure, the domains being small regions throughout which, except near the boundaries, the magnetization is unidirectional. In an unmagnetized iron crystal the domains may be supposed to be magnetized along one or other of the six equivalent easy directions along the cubic axes, equal numbers of domains, or, more precisely, numbers giving equal aggregate volumes, being magnetized along each of the six directions. In comparatively small fields, by a mechanism to be further considered, the directions change to the easy directions which are nearest to the field direction. On the completion of this change, for the field along the [100], [110] and [111] directions, the magnetization of the domains will be along the one, two and three nearest easy axes, giving a resultant magnetization in the field direction of I_0, $I_0/\sqrt{2}$ and $I_0/\sqrt{3}$ respectively, in reasonable agreement with the values at which the sharp bends occur in the experimental curves (figure 2.6). Further increase in the magnetization in the [110] and [111] directions can occur only by rotation of the direction of the domain magnetization from an easy direction towards that of the field. The rotational processes, involving continuous change in direction of magnetization as the field increases, overlap the processes involving abrupt changes of direction in the lower fields, but are predominant above the bends in the magnetization curves.

The complicated interrelations between the magnitudes and directions of magnetization and field found experimentally for single crystals in the "rotational range" can be formally coordinated in a remarkably satisfactory manner on the basis of energy and symmetry considerations. The energy associated with the magnetization of a crystal clearly depends on the direction of the magnetization with respect to the crystal axes, and will be expressible as a power series in the direction cosines. For a cubic crystal, let α_1, α_2 and α_3 be the direction cosines of I_0 referred to a coordinate system defined by the cubic axes. Then it is not difficult to show that to the lowest order term which is neither zero nor constant the dependence of the free energy per unit volume on the direction of magnetization is given by

$$F_c = K_4(\alpha_2^2\alpha_3^2 + \alpha_3^2\alpha_1^2 + \alpha_1^2\alpha_2^2). \qquad \ldots\ldots(2.17)$$

This term alone is sufficient to cover the experimental results reasonably well for both nickel and iron, though for iron a slight improvement is obtained by including a sixth order term, with a further coefficient K_6. The values of the anisotropy coefficients, K, are to be determined from the experimental results. Perhaps the simplest method is from the difference in the work required to magnetize a crystal in different directions, readily obtained from magnetization curves such as those in figure 2.6. From (2.17), the difference for the [110] ($\alpha_1 = \alpha_2 = 1/\sqrt{2}$; $\alpha_3 = 0$) and the [100] ($\alpha_1 = 1$, $\alpha_2 = \alpha_3 = 0$) directions, for example, is $K_4/4$. By adding to F_c

in (2.17) an expression for the external free energy in a field, dependent on the relative directions of H and I_0, and finding the conditions for minimizing the total free energy, theoretical curves may be calculated for the variation of the parallel and transverse components of the magnetization with the magnitude and direction of the field, the K values being chosen to give the best fit with the experimental curves. The essential test of the value of the treatment is that different methods of evaluating the crystal anisotropy coefficients from the data lead to consistent results. For iron (body-centred cubic) the approximate value of K_4 is $+4{\cdot}0\times10^5$ erg.cm^{-3} For nickel (face-centred cubic), the value is much smaller, and negative (the easy directions are the cube diagonals): $K_4=-0{\cdot}5\times10^5$ erg.cm^{-3} For cobalt, which is hexagonal, the expression for F_c begins with the quadratic term, and may be written in the form

$$F_c=K_2\sin^2\phi+K_4\sin^4\phi, \qquad \ldots\ldots(2.18)$$

where ϕ is the angle between I_0 and the hexagonal axis. Magnetization is much easier along this axis than at right angles to it, and correspondingly the coefficient K_2 is positive and large:

$$K_2=4\times10^6\ \text{erg.cm}^{-3}.$$

Measurements on magnetostriction have also been made for a number of single crystals. For cubic crystals the important coefficients are λ_{100} and λ_{111} giving the relative change of length for change of magnetization from 0 to I_0 in the [100] and [111] directions. For nickel, both λ_{100} and λ_{111} are negative, but the magnitudes are different: $\lambda_{100}=-5\times10^{-5}$, $\lambda_{111}=-3\times10^{-5}$. For iron, λ_{100} is positive ($1{\cdot}8\times10^{-5}$) and λ_{111} negative ($-1{\cdot}3\times10^{-5}$). The magnetostrictive properties of single crystals can be treated in a similar way to the magnetic properties, and indeed, through the work of Akulov, Becker, Gans and others, the elastic, magnetic and magneto-elastic properties of "ideal" crystals can be satisfactorily embodied in a general formal scheme. The extent of the coordination effected for cubic crystals by a scheme involving, apart from the elastic coefficients, only a single anisotropy coefficient, K, and two magnetostriction coefficients, λ_{100} and λ_{111}, is very remarkable.

The details of the formal scheme for single crystals are of rather specialized interest, but the possibility of "summarizing" a whole complex of observational data in the numerical values of two or three coefficients is clearly of great importance. It is the task of atomic theory to account for the magnitude of these coefficients (in the same way that it attempts to account for an experimentally determined value of χ, say). Quasi-magnetic interactions between the elementary carriers of the magnetic moment have been extensively investigated, but it is now clear that they will not explain the magnitude, or in some cases even the sign, of the anisotropy and magnetostriction coefficients; it seems probable that spin-orbit coupling effects are involved, but the problem is a very complex one, unlikely to be dealt with quantitatively in a satisfactory way until the component parts are more clearly sorted out and better understood.

As ferromagnetic materials, single crystals are hardly common, and some of the general relations between stress, strain and magnetization are better taken up in connection with materials of a more usual kind. One particular simple point may, however, be mentioned here, namely that a crystal under stress is effectively a crystal with symmetry modified to an extent depending on the strain set up.

Thus a cubic crystal, with six equivalent easy directions, and isotropic positive magnetostriction, when stretched acquires an easy direction of magnetization along the axis of tension. Ultimately the natural crystal anisotropy may become negligible compared with the strain anisotropy. A precise quantitative treatment of the effects of stresses on polycrystalline material is possible only on the basis of a knowledge of the values of the anisotropy and magnetostriction coefficients of single crystals, so that single crystal measurements have a very practical value quite apart from their fundamental importance for theory.

2(ix). *Magnetization curves, strains and domains*

When a magnetic field is applied to a specimen in an initially demagnetized state, the magnetization at first increases slowly, and mainly reversibly, with the field, then more rapidly, and mainly irreversibly, and then slowly, and mainly reversibly again towards a fairly well defined saturation value. (See figure 2.3.) In the third stage the main elementary process is rotation of the magnetization of a domain from an easy direction towards the direction of the field, the direction taken up in any field corresponding to an equilibrium between the applied field and the forces associated with the natural crystal or the strain anisotropy of the particular domain. It should be stressed that this is not the only process which occurs during the third stage of magnetization. Near saturation, the variation of magnetization with field may usually be represented empirically by

$$I/I_0 = 1 - \alpha/H^2 - \beta/H + \gamma H. \qquad \ldots\ldots (2.19)$$

The anisotropy effect can give rise only to the term in $1/H^2$, which is usually predominant if the fields are moderate. The term in H may be predominant in high fields, but it is always small, and is ordinarily taken to correspond to an increase in the "true" magnetization, mentioned in connection with the magneto-caloric effect. The term in $1/H$ has been established beyond doubt in many cases, but its origin is uncertain; it could be accounted for as a residual effect arising from the presence of domains containing a relatively small number of atoms. In the second stage of magnetization, the main elementary process is the reversal, or change through a finite angle of the direction of magnetization of a domain (that is the effectively simultaneous change of orientation of a large number of spins). In the first stage, although the change with field is reversible, the magnitude is too great to be accounted for, in general, by the rotation process; the main process is a reversible shift of domain boundaries, so that domains magnetized with a component in the field direction "grow" at the expense of their neighbours. The experimental evidence behind these statements about the processes occurring in the first two stages of magnetization must now be briefly indicated.

The discontinuous character of the process of magnetization, rendered apparent to the ear in the now well known and often repeated Barkhausen experiment, has been extensively investigated with a view to obtaining reliable estimates of the sizes of the discontinuities; or, since the discontinuities may be presumed to correspond to reversals or abrupt changes of direction of the electron spins over small regions of unidirectional magnetization, of the sizes of domains. Among the most comprehensive studies of the effect in ordinary unstrained material are those of Bozorth and Dillinger (1929–32), who developed an ingenious method which enabled the variation of the average volume of the regions of reversal

along the magnetization curve to be determined. A typical curve is shown in figure 2.7. It was found that almost the whole of the change of magnetization on

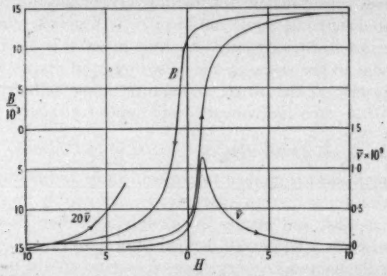


Figure 2.7. The Barkhausen effect for annealed iron. The $\bar{v}$ curve shows the average volume of material (cm³) over which the magnetization reverses direction in an individual Barkhausen jump at the corresponding part of the hysteresis curve. (*After Bozorth and Dillinger*, 1930.)

the steeper part of a hysteresis loop occurred by the reversal in groups corresponding to not less than 10^{10} atoms. The average size is greatest at or near the steepest part of the loop, where, among the materials examined, it ranged from about $1{\cdot}2 \times 10^{-9}$ cm³ (about 10^{14} atoms) for annealed iron to about $4{\cdot}5 \times 10^{-9}$ cm³ (about 4×10^{15} atoms) for 50% ferro-nickel. The maximum size is not obviously related to the crystal grain size; the sizes found were in fact of the same order in single-crystal as in polycrystalline material. The order of size is that of the separate crystals in many well annealed materials, but considerably larger than that in hard-worked samples.

The Barkhausen discontinuities, though difficult to account for in quantitative detail, confirm the existence of a domain structure for which the earlier evidence was much more indirect. There is further confirmation in the patterns, discovered by Bitter (1931), which are obtained with fine colloidal suspensions of ferromagnetic oxides on smooth polished surfaces of ferromagnetics. Though the detailed interpretation of these patterns is still obscure, they certainly appear to give direct visual evidence of a microscopic magnetic structure on the same scale as is indicated by other types of investigation.

Although ordinary Barkhausen discontinuities are so small, it is possible, by the use of suitably strained material, to obtain discontinuities corresponding to a large fraction of the change in magnetization for complete reversal from $+I_0$ to $-I_0$. The experimental study of large Barkhausen discontinuities initiated by Sixtus and Tonks in 1931 has thrown much light on the character of the magnetization process. The conditions under which large discontinuities may occur are perhaps most simply approached by considering a general relation for the effect of stress on magnetization. In so far as the energy of a ferromagnetic wire is determined uniquely by the field, H, or the magnetization, I, and the length, l, or the tension, Z, the following reciprocal relation holds:

$$\frac{\partial(\delta l/l)}{\partial H} = \frac{\partial I}{\partial Z}. \qquad \ldots\ldots(2.20)$$

Positive magnetostriction implies an increase of magnetization with tension, or a decrease with compression, and conversely for negative magnetostriction. There are admirable illustrations of these effects in the curves obtained by Ewing over fifty years ago, and given in his book, for iron and nickel under tension and compression. Since the revival of interest in this field in the late twenties, many interesting sets of curves have been published, of which an example is shown in figure 2.8. When the crystal structure is cubic it is not difficult to show, from

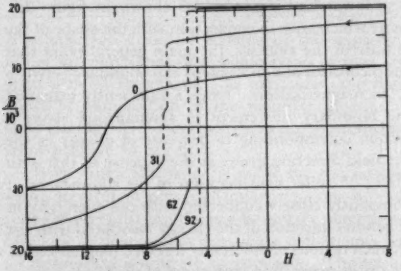


Figure 2.8. Effect of tension on the hysteresis curves of material with positive magnetostriction. Ferro-nickel, 15% Ni. The numbers on the curves give the tension in kg.mm^{-2}. (*After Sixtus and Tonks*, 1931.)

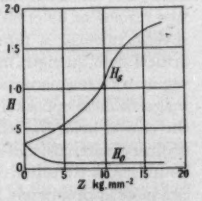


Figure 2.9. Limiting field, H_0 for the propagation of reversed magnetization, and starting field, H_s, for permalloy under tension, Z. (*After Preisach*, 1932.)

(2.17) and similar equations for the magneto-elastic energy, that the effect of the artificial anisotropy produced by tension becomes comparable with that of the natural anisotropy when Z is of the order $K/3\lambda$ (which gives values of about 4 kg. mm^{-2} for nickel, and some twenty times greater for iron). With a polycrystalline wire of material of negative magnetostriction, tension tends to develop easy directions at right angles to the axis of the wire; ultimately the spontaneous magnetization of the domains is entirely in these directions, and magnetization along the length of the wire takes place entirely by the reversible rotation process. The formula derived for the initial susceptibility under these conditions ($\kappa_0 = (2/9)I_0^2/\lambda Z$, where Z is the applied stress) gives values in excellent agreement with those found experimentally. For material of positive magnetostriction (such as that for which curves are given in figure 2.8) the spontaneous magnetization of the domains is along the length of the wire. The magnetization curves show that when the wire has been magnetized to saturation in one direction, application of a gradually increasing field in the opposite direction may, at a critical value, result in the almost complete reversal of the magnetization in a single jump. It is reasonable to suppose that reversal starts at one or more reverse "nuclei", and that the boundaries of the regions of reversed magnetization travel along the wire with a finite velocity. In the experiments of Sixtus and Tonks steady reverse fields, H, were applied, and reversal was started near one end of the wire by a larger localized "starting field", H_s. The velocity of propagation was measured by the aid of thyratron circuits controlled by the impulses set up in

suitably placed search coils by the passage of the boundary of the reversed magnetization region. For a given tension, it was found that above a critical value, H_0, of the reverse field, the velocity of propagation of the reversed magnetization along the wire was approximately proportional to $H-H_0$, the proportionality factor varying little with tension, and H_0 decreasing as the tension increased. Later work by Preisach shows that the minimum starting field, H_s, and H_0 tend to the same limit as the tension is decreased, and that with increasing tension H_s increases, and H_0 decreases to a lower limiting value, as shown in figure 2.9. Leaving aside the special problems which arise in connection with the shape of the boundary and the actual magnitude of the velocity, the main general point that emerges is that a Barkhausen jump consists of a movement of a boundary between two regions of differently directed magnetization. Once a sufficiently extended reverse nucleus is formed, the boundary movement is spontaneous above a critical field strength in a direction corresponding to decrease of energy as the region with magnetization in the field direction grows at the expense of that with magnetization oppositely directed. A fuller consideration of the significance of the critical field, H_0, which is obviously closely connected with the coercivity in ordinary materials, and also the general question of the energy associated with the boundary, or transition zone, between regions of differently directed magnetization can more appropriately be taken up in connection with some of the later work on magnetization curves, which it is hoped will be discussed in a subsequent report.

The interpretation of the initial stages of magnetization as associated with reversible movements of domain boundaries is mainly due to Becker, who, on this basis, in 1930 derived a relation between the initial susceptibility and the internal strains in the material. The treatment, though admittedly of a very tentative kind, and certainly incorrect in detail, led to results in surprisingly good agreement with experiment, and made clear certain essential factors controlling the magnetization process. The internal stresses are regarded as alternating tensions and compressions, and along a line in the direction of an applied field the stress is idealized as sinusoidal, with amplitude Z_1. When the field is zero, alternate regions under tension and compression will be magnetized at right angles to each other, in such a way that the resultant magnetization is zero, the domain boundaries (in this case 90° walls) having their central points or planes where the stress is zero. With a field, H, the boundaries move so as to increase the volume of those domains magnetized in the field direction, until the rate of decrease of the magnetic energy proportional to HI_0 is equal to the rate of increase of the magneto-elastic energy proportional to $Z\lambda$. If the shift is small, it is readily found that the magnetization in the field direction is given by

$$I_H = HI_0^2/\pi\lambda Z_1,$$

independent of the period owing to the special assumption of a sinusoidal distribution of stress. (With a given Z_1, the smaller the period the less the gradient of Z at a boundary, and hence the greater the shift for a given H; this is balanced by the number of boundaries being correspondingly smaller.) The expression for the initial susceptibility is, therefore,

$$\kappa_0 = I_0^2/\pi\lambda Z_1, \qquad \ldots\ldots(2.21)$$

indicating that the initial susceptibility will be smaller the greater the variations in

the internal stress. In a general way this semi-theoretical conclusion is in accordance with what is suggested by experience. When internal strains are as far as possible removed by suitable heat treatment, there remains, even in the purest material, a residual magnetostrictive strain set up as the material cools down from the Curie point, the associated stress being of the order $E\lambda$, where E is Young's modulus. Thus the order of magnitude of the maximum attainable initial susceptibility is given by

$$(\kappa_0)_{max} = OI_0^2/3\lambda^2 E. \qquad \ldots\ldots (2.22)$$

A comparison made by Kersten of this expression with the highest measured initial susceptibilities of a whole series of iron nickel alloys showed remarkably good general agreement. In particular, the high susceptibilities attainable in alloys of composition in the neighbourhood of 78% nickel become immediately intelligible, for in this region the mean magnetostriction coefficient changes from positive to negative.

Even when allowance is made for the necessary brevity of statement, this short account of proposed interpretations of ordinary magnetization curves in terms of domains and domain boundary movements may well give the impression of a lack of definiteness just where definiteness would be most desirable. Such an impression is entirely appropriate. The background to the work since 1934 on ferromagnetism includes many things to which no reference has been made, but it does not include a clear understanding of the status of domains; and, although some of the later work adds definiteness to various aspects of the general scheme, it cannot be said that the period up to 1946 has yielded a fully satisfying answer to all the fundamental questions appreciated as outstanding in 1934.

§ 3. INTRINSIC MAGNETIZATION. THEORETICAL

3 (i). *Introductory*

The fundamental problem of ferromagnetism is that of the intrinsic magnetization and of its variation with field and temperature. The variation with field, however, is usually of subsidiary interest. Above the Curie point region, the spontaneous magnetization, that is the intrinsic magnetization in zero field, is zero, and the intrinsic magnetization may be identified with the apparent magnetization, which, with ordinarily attainable fields, varies linearly with the field, except in the immediate neighbourhood of the Curie point. Above the Curie point, therefore, it is the susceptibility and its temperature variation which is to be considered. Below and not too near the Curie point, the change with field of the intrinsic magnetization is small even with high fields. Interest centres in the temperature variation of the spontaneous magnetization. The variation with field of the apparent magnetization, which depends on a number of "secondary" factors (§ 2 (viii), (ix)), comes into consideration only in so far as it may be involved in making estimates of the spontaneous magnetization. In the lower temperature range the spontaneous magnetization can be taken, for most practical purposes, as equal to the quasi-saturation magnetization, as measured in fields of a few thousand oersted. Much more precise estimates can be made for the whole temperature range from magnetization and magneto-caloric data, using the extrapolation procedures developed by Weiss and his collaborators.

The coordination of the experimental facts by the Weiss molecular field hypothesis (§2(iii), (vii)), quantitatively successful to a remarkable degree in spite of its many imperfections in detail, and Heisenberg's interpretation of the molecular field as arising from quantum mechanical interchange interaction effects (§2(vi), (vii)), have already been described. Subsequent general theoretical work will be outlined in this section, various special aspects of it being left for consideration in connection with experimental results in §4. The general work is of two main types. In one, the attempt is made to improve the Heisenberg treatment, the quantitative failure of which has been emphasized, approaching the problem in essentially the same way, that is by adapting the Heitler-London method as used for the hydrogen molecule. In the other, a different approach is adopted, namely that which has been so successful in the general theory of metals, the effective electrons being treated as "itinerant", or shared collectively, through the metal, rather than as being associated with particular atoms. In neither of the two methods of attack on the theory of ferromagnetism is an attempt made at a rigorous solution of the problem in all its complexity. Rather, a simplified problem is treated, that of determining the behaviour, under certain conditions, of an "idealized" metal. The idealizations being in roughly opposite directions, the two methods are complementary rather than antagonistic, and bring out different features of interest. It is almost impossible to convey briefly a proper impression of the first type of work, in which the centre of interest is less in new physical points of view introduced than in the mathematical technique employed to deal with a problem which remains of great complexity even when highly idealized physically; and little more than a bare indication of some of the contributions in this field is given. The collective electron treatment does introduce new physical points of view, and it has the great advantage of linking up the treatment of the ferromagnetic metals with that of metals generally, so that a rather fuller survey would seem appropriate.

3(ii). *The Heitler-London approach*

Shortly after Heisenberg's 1928 paper it was shown by Bloch (1930) that a rigorous solution of the equations could be obtained for low temperatures when all but a few of the electron spins were pointing in one direction. In these circumstances the number of arrangements of the reversed spins, and the associated energies of the crystal can be calculated, assuming effective interchange interaction only between nearest neighbours. The name "spin wave" given to Bloch's treatment derives from the form of the functions representing admissible solutions to the Schrödinger equation for the problem. A partition function is constructed in the usual way, and from it can be derived the dependence of magnetization on temperature, the result, valid only near absolute zero, where saturation is nearly complete, being

$$I = I_0\{1 - A(kT/J_0)^{3/2}\}, \qquad \ldots\ldots(3.1)$$

where J_0 is the interchange interaction integral and A a numerical constant depending on the form of the crystal lattice and of the order 0·1. The experimental results for the variation of I with T at low temperatures are not in agreement with the exponential type of law given by the simple modified molecular field theory (see figure 2.4), but are fairly well represented by a T^2 law; the very careful

work of Fallot (1936), however, on iron, nickel and an iron-aluminium alloy, shows that, if due weight is given to the measurements below 90° K., more satisfactory agreement is obtained with a $T^{3/2}$ law, so that, to this extent, there is confirmation of the Bloch formula (3.1). The coefficient of $T^{3/2}$ in (3.1) does not appear to have been critically considered in relation to the experimental results.

A method of interpolating between the Bloch results, valid at low temperatures, and the Heisenberg results, for temperatures in the Curie point region has been given by Ludloff (1934), and further developed and extended to the case where there is more than one electron per atom, by Oertel (1937). The general agreement with experiment is poor. Contributions to the technical problem have been made by Kramers (1936), and Opechowski (1937). The latter finds that a more complete series expression for the magnetization at low temperatures as a function of temperature includes a term in T^2 as well as in $T^{3/2}$. Above the Curie temperature, the inverse of the susceptibility is shown to be theoretically expressible as proportional to $\{T-\theta+\Sigma a_n/T^n\}$, and the coefficients a_1, a_2 and a_3 are computed. As far as the T^{-1} term the formula is equivalent to that of Heisenberg, which agrees less well with experiment than without the T^{-1} term. Inclusion of the T^{-2} term leads to appreciable improvement, but after the very troublesome calculation of the coefficients a_3, the inclusion of the T^{-3} term is found to lead to results which, from an experimental standpoint, are hopeless. Series calculations undertaken without *a priori* grounds for expecting rapid convergence often prove disappointing in this way.

The more obviously unsatisfactory features of Heisenberg's detailed results undoubtedly arise in large part from the virtual impossibility of making an accurate estimate of the energy distribution of states of given total spin. The assumption of a Gaussian distribution leads to far worse agreement with experiment than the simple treatment, in which, in effect, the average energy only is considered. Fay (1935) developed a method which enabled him, in a simple case, to take half the interaction into account rigorously. The crystal was regarded as built up of units of eight atoms at the corners of a cube. The energy levels of a unit of this kind, eight electrons in similar orbits at the corners of a cube, had already been worked out. On the nearest neighbour assumption, each atom interacts with three neighbours in its own cube, and three in other cubes. The second set of interactions is dealt with by the average interaction energy process, the first set by taking into account the detailed energy level structure of the cube states. The net result is that the Curie temperature is related to the interaction integral by a proportionality factor some 15% smaller than in the simple theory, but the reduced curves for the magnetization and susceptibility against T/θ are almost identical with those shown in figure 2.4, except for a slight curvature in the $1/\chi$, T curve above the Curie point. This suggests that a fully rigorous treatment of the interchange effects, could it be carried out, might give far less difference from the results of the simple treatment than did the approximation method used by Heisenberg.

The close analogies between the change from order of electron spins in the ferromagnetic state to disorder in the paramagnetic state and the change from order to disorder of atoms of two kinds in alloys, suggest that the concepts of short range and long range order, and the mathematical methods developed by Bethe,

Peierls and others for dealing with it in alloys, might profitably be taken over in the treatment of ferromagnetism. The most comprehensive work along these lines seems to be that of P. R. Weiss, some results of which have been reported by Van Vleck (1945), whose review of theories of ferromagnetism should be consulted in this and other connections. The approach of Weiss is not dissimilar to that of Fay, but the method is much more general; the results reported indicate that the conditions obtained for the occurrence of ferromagnetism as dependent on the number of nearest neighbours in the various types of lattice are in agreement with those obtainable in a few simple cases (e.g. the linear chain) by exact methods, and not incompatible with experiment; and that a nearly linear relation between $1/\chi$ and T is obtained above the Curie point.

3 (iii). *Fluctuations*

Much of the trouble in detail in the theory of ferromagnetism arises from the assumption that interchange interaction is negligible except between nearest neighbours, an assumption which seems not unreasonable in view of the presumably rapid exponential decrease in the value of the interchange interaction integral, after the maximum is passed, with increase of internuclear distance for two interacting atoms. If the number of effectively interacting neighbours is small, it is not legitimate to substitute the average value of the relative magneti-

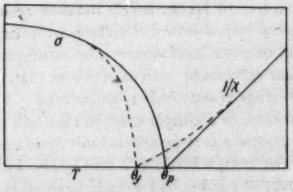


Figure 3.1. Effect of fluctuations of the molecular field on the spontaneous magnetization and the susceptibility of ferromagnetics. Full curves, simple theory; broken curves, theory incorporating effects of fluctuations. (After Néel, 1940.)

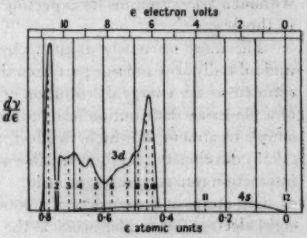


Figure 3.2. Energy distribution of states for copper. The number of levels per unit energy range, $d\nu/d\epsilon$, is plotted against the energy, ϵ, in atomic units, with the lowest energies to the left. The 3d and 4s bands are shown separately. Vertical lines indicate the portion of the bands filled by 1, 2, . . . 12 electrons, respectively. (After Slater, 1936.) The scale at the top gives the energy in electron volts.

zation of the specimen as a whole for the instantaneous value of the relative magnetization of any small group of atoms, as is done, for example, in passing from (2.13) to (2.14) in §2(vii). The statement made earlier that the energy distribution of states of given total spin should be taken into account and not merely the average energy is perhaps a somewhat sophisticated expression of the situation; an alternative statement is that it is necessary to take into account the effect of "fluctuations" (e.g. in the local intensity of magnetization). From what has been

said it will be clear that fluctuations would be taken into account automatically in a rigorous treatment, but in view of the only too obvious difficulties of this, a separate consideration of fluctuations is well worth undertaking. Very comprehensive studies have been made by Néel (1932, 1934, 1940), who obtains many illuminating results. He regards the long range forces as covered by a Weiss molecular field treatment, and deals in detail with the statistics of the short range effects. For simplicity, but also in harmony with the physical situation, he treats a "carrier" as having only two possible settings, and a pair of "nearest neighbours" as having two possible values of the mutual potential energy according as the moments are parallel or anti-parallel. Taking the number of nearest neighbours as $2p$, exact formulae are obtained for $p=1$ (the results of Ising, 1925, for a linear chain were obtained independently) and $p=2$, and by ingenious mathematical devices, approximate formulae are obtained for any value of p. The character of the deviations from the results of the simple molecular field theory (figure 2.4) is shown diagrammatically in figure 3.1. The deviations are greater the smaller the number of interacting neighbours. Taking the interactions as equal between any pair of atoms in a finite block of N atoms, Néel finds that the experimental results for nickel require a value of N of about 750. The noteworthy point about Néel's treatment of fluctuations is that the divergences between the simple theory and experiment in the neighbourhood of the Curie point in regard to the spontaneous magnetization below, the susceptibility above, and also the tail in the specific heat curve, mentioned in connection with figure 2.5, are all covered reasonably well by the same value of the fluctuation parameter N. No great significance can perhaps be attached to the actual value of N found; it does, however, suggest that interchange interaction effects may be appreciable between pairs of atoms much more widely separated than nearest neighbours.

3 (iv). *The collective electron approach*

All the earlier atomic theories of ferromagnetism are confronted with the difficulty that the saturation magnetization does not correspond to an integral number of Bohr magnetons per atom, a difficulty which could be met only by an artificial and *ad hoc* conception of a metal as containing an equilibrium mixture of neutral atoms and ions of different moments, including zero. It was suggested by Stoner (1933) that, with the electron energy band treatment of solids, non-integral values of the magnetic moment per atom in both elements and alloys might be accounted for as arising from the equilibrium set up when the tendency to anti-parallel setting of the electron spins, normally giving the state of lowest energy, is opposed by the tendency to parallel setting, due to the interaction integral being positive. Little was then known about the form of the electronic energy bands in the transition metals, and the treatment was necessarily qualitative. Considerably greater precision became possible on the basis of a suggestion by Mott (1935) on the form of the energy bands in the ferromagnetic metals. In nickel it was suggested that a narrow d band (that is the band in the metal arising from the states of the d electrons in the separated atoms) was overlapped by a much wider s band, and that the top of the Fermi distribution came at a point corresponding to about 0·6 s electrons per atom in the s band, with a deficit of the same number in the d band. For reasons discussed earlier (§ 2 (vi)) ferromagnetism is to be attributed to

the interactions of d electrons rather than of s electrons. The full complement of electrons in the d band is 10 per atom, with 5 spins pointing in each direction. "Holes" in a band are to a large extent magnetically equivalent to the same number of electrons in an otherwise empty band (see Stoner, 1936 a). With 9·4 electrons in the band, it could be provisionally assumed that owing to positive interchange interaction, the spins of as many as possible, namely 5, would be aligned parallel, leaving an excess of 0·6 parallel spins per atom. The number is of course primarily chosen to give agreement with the low temperature saturation magnetization, but Mott shows that it is in harmony with other experimental facts.

The first attempt at a more quantitative treatment of ferromagnetism on a collective electron basis is that of Slater, in two papers (1936) on the ferromagnetism of nickel. The form of the electron energy bands, as estimated by Slater from the calculations of Krutter (1935) on copper, made by the Wigner-Seitz method, is shown in figure 3.2. In copper there are 11 electrons, in addition to the 18 electrons in 1s, 2s, 2p, 3s and 3p states which would give completely full separated bands, high and very narrow, well to the left of those shown in the diagram. The 3d band is full, and the conductivity and other electronic properties of the metal are due to electrons at the top of the distribution in the half filled 4s band. In nickel, assuming that the form of the bands is much the same, the distribution indicated in the figure for 10 electrons, with about 0·5 electrons in the s band, and the same number of holes in the d band, is similar to that suggested by Mott. No great weight should be attached to all the irregularities in the d band curve, but it should perhaps be underlined that the curve, though it may well be incorrect in detail, is not speculative, but is based on elaborate numerical computations. The distribution obtained, with the sharply defined upper limit for the electron energy, applies to the absolute zero of temperature, and, since interchange interaction is neglected, each state is either occupied by two electrons of opposite spins, or is empty. Any re-orientation of spins to give an excess in one direction necessarily involves an increase of energy. All the states of + and − spin below the energy value ϵ_0, say, are fully occupied, and a change of orientation of a spin can take place only if there is an increase in the electron energy from a value below to a value above ϵ_0. The energy increase for a given increase in the number of parallel spins can be calculated if the band form in the neighbourhood of ϵ_0 is known. With a positive interchange interaction, there is a decrease in energy with increase in the number of parallel spins, and this may more than compensate the increase in the Fermi energy. Slater makes a theoretical estimate, admittedly rough, of the interchange interaction integrals in a manner appropriate to the collective electron formulation of the problem, and he shows that the interchange energy increases (numerically) more rapidly than the Fermi energy with increase in the net number of parallel spins per atom at least over a sufficient range to ensure the parallel alignment of most of the available spins in the d band in nickel at absolute zero. The energy associated with this magnetization, moreover, is in fair agreement, as to order of magnitude, with that estimated from the observed Curie temperature. In the second paper, by the application of Fermi statistics to the electrons in the d band, Slater investigates the temperature variation of the ferromagnetism of nickel. The method used is not well adapted to a precise determination of the magnetization as a function of the temperature,

particularly when the energy distribution of states in the band is available only in the form of a numerical table. To within the wide errors of calculation, however, there is agreement with experiment.

The statistical problem connected with the collective electron ferromagnetism may be stated, in its simplest form, as follows: an electron energy band, of given form, contains N electrons, which at absolute zero, and in the absence of positive interchange interaction, occupy the band in balanced pairs up to an energy ϵ_0; it is required to find the dependence of the distribution (and hence of the magnetization, and the energy) on the magnitude of the interchange interaction, on the applied field, and on the temperature. This problem has been treated in detail by Stoner (1938 a, 39 a) for a band of standard form (in which the energy density of states is proportional to the square root of the energy), and on the assumption, phenomenologically justified, and theoretically at least as a first approximation, that the interchange interaction effects give rise to a term in the energy expression proportional to the square of the magnetization. Of the "manageable" band forms, there can be no question but that the one chosen for detailed investigation is of the greatest general interest, both for theoretical reasons, and for its wide practical applicability as an approximation over relevant ranges to actual band forms in metals. The electrons must be treated on the basis of Fermi-Dirac statistics (see Stoner, 1939 b), and the necessary precision in the numerical work was made possible only by the use of previously computed tables of Fermi-Dirac functions (McDougall and Stoner, 1938). It is convenient to express the resultant magnetic moment, M, as a fraction, ζ, of the maximum moment of the N effective electrons, so that $\zeta = M/N\mu$. (It is to be noted that N will not ordinarily correspond to an integral number of spins per atom; it is the effective number of electrons, or holes, in the band concerned, and for nickel would be expected to correspond to about 0·6 per atom.) As a measure of the interchange interaction, $k\theta'$ is used, and is such that $2k\theta'\zeta$ is the difference in energy for a spin parallel and anti-parallel to the magnetization (cf. (2.10) and (2.14)). Finally, it is convenient to use a reduced temperature, kT/ϵ_0, where ϵ_0 is the maximum electron energy, measured from the bottom of the band, when the electrons are in balanced pairs in the lowest energy states in the band. (For nearly full bands, such as the d band for nickel in figure 3.2, "holes" would be considered, and the energy measured from the top of the band.) The results of the calculations may be approached by way of those previously obtained (Stoner, 1936 a) for the effect of positive interchange interaction on electron spin paramagnetism, illustrated in figure 3.3. The character of the $1/\chi$, T curves corresponding to (1) depends on the magnitude of ϵ_0. It will be readily seen that in the limit $\epsilon_0 \to 0$, corresponding to classical statistics, the relation is that for a Curie law paramagnetic. If, however, ϵ_0 is large (and for strictly free electrons in a metal, kT/ϵ_0 would be of the order $10^{-5}T$), the relation corresponds to a feeble paramagnetism, practically independent of temperature over the experimental temperature range, such as is found for many metals. With positive interchange interaction the susceptibility is increased, in a manner corresponding, exactly as in the Weiss treatment, to a constant decrease in $1/\chi$. In the classical limit the Curie temperature θ is identical with θ'. With the Fermi-Dirac treatment it is lower; and unless the interchange interaction is large enough, $k\theta'/\epsilon_0$ exceeding $\frac{2}{3}$ for the parabolic band considered, it will not result in ferro-

magnetism. It follows that, with feasible values of the interchange interaction, the effective width of the relevant part of the unfilled energy band must be small, so that, apart from any other considerations, ferromagnetism would not arise from the s band in figure 3.2.

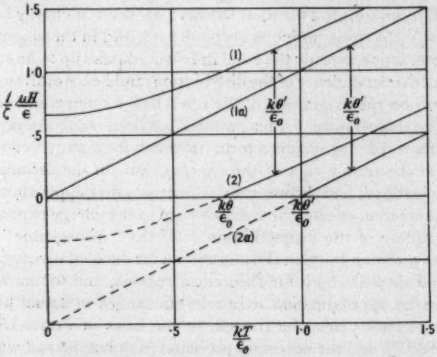


Figure 3.3. Effect of interchange interaction on the electron spin paramagnetism. ζ, relative magnetization ($M/N\mu$); μ, Bohr magneton; ϵ_0, Fermi energy; θ', interaction parameter. The susceptibility per electron is given by $\chi_0=\zeta\mu/H$. A quantity proportional to the inverse of the susceptibility is plotted against the reduced temperature: (1), with zero interchange interaction; (2), with positive interaction. The straight lines (1a) and (2a) are the asymptotes to (1) and (2), and correspond to the results which would be obtained with classical statistics.

The numerical results for the spontaneous magnetization below and the paramagnetism above the Curie point may be obtained, after considerable labour, in such a form that they can be plotted in the same way as in figure 2.4, which is probably the most convenient for purposes of comparison with experiment. Since there is an additional parameter, a set of curves instead of a single curve is required. A number of curves, for different values of $k\theta'/\epsilon_0$, are shown in figure 3.4. It will be noticed that ζ/ζ_0 is here used rather than ζ as in figure 2.4. This is because in the $k\theta'/\epsilon_0$ range between $\frac{2}{3}$ (the lower limit for spontaneous magnetization) and $2^{-1/3}$ ($=0{\cdot}794$), the equilibrium at absolute zero corresponds to an incomplete alignment of the electron spins, ζ_0 increasing at first rapidly and then more slowly from 0 to 1. There are a number of striking points about these curves. The spontaneous magnetization curves are below the classical curve, but for the whole range of $k\theta'/\epsilon_0$ values do not differ greatly from it. The inverse susceptibility curves, however, are widely different, particularly for the lower values of $k\theta'/\epsilon_0$, where ζ_0 is less than 1. For the higher values the curves are almost linear over the range shown, but the apparent slopes even there differ considerably from the slopes, equal to ζ_0, of the asymptotes. (For the curve labelled 0·680, $\zeta_0=0{\cdot}5$; for the other curves $\zeta_0=1$.) The reason for some, at least, of the inconsistencies in the values calculated for atomic moments (or numbers of effective electron spins per atom) from low and high temperature measurements by the older methods immediately becomes apparent.

In the second paper (1939a) the temperature dependence of the electronic energy and specific heat is considered. The electronic specific heat at very low temperatures, the "magnetic" excess specific heat rising to a maximum at the Curie point, and the excess specific heat above the Curie point are covered by a single scheme, and, with the parabolic band form, are all completely determined

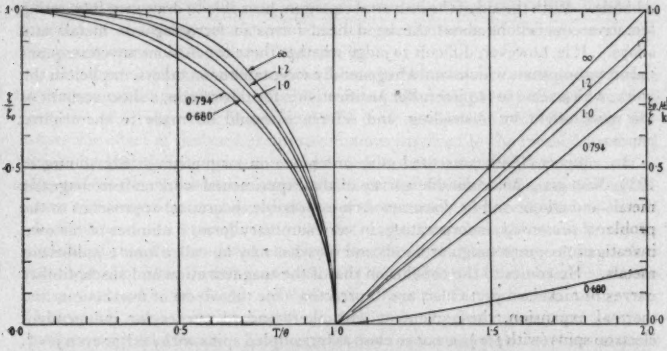


Figure 3.4. Curves showing reduced magnetization below and reduced inverse susceptibility above the Curie point as a function of T/θ. ζ, relative magnetization; ζ_0, relative magnetization at absolute zero. The numbers on the curves give the values of $k\theta'/\epsilon_0$. The limiting curves labelled ∞ correspond to the use of classical statistics. (Stoner, 1938a.) (*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

by the three parameters N, ϵ_0 and θ'. The general character of the results is completely in accordance with experiment, but detailed comparisons have not yet been made. It may be said at once, however, that agreement in detail with the whole set of magnetic and thermal experimental data is certainly not obtained; and added, that if it were, it could not be other than fortuitous. This for a number of reasons. First, there are a number of factors, such as fluctuations, which are not embodied in the treatment; second, the band form assumed cannot be better than a rough approximation to that in any actual ferromagnetic, and a fit over the whole temperature range with a constant value of ϵ_0 is even less likely than a fit of lattice specific heat data over the whole range with a constant Debye θ_D; third, both ϵ_0 and θ' depend on the lattice spacing, the variation of which with temperature (the importance of which is stressed by Néel in a related connection) is not considered; and finally, with band overlap, there may be a non-negligible variation with temperature of the effective number of electrons. The treatment, in short, gives a comprehensive survey of the interlocked thermal and magnetic properties of a "model" of a ferromagnetic based on the simplest possible assumptions which are consistent with the general theory of the metallic state. Even so, the details are surprisingly complicated. Only when the implications of the simplest physically reasonable assumptions about the electronic structure of a ferromagnetic metal are known with some precision, and the experimental facts are critically considered in the light of them, can the need for, and the effect of refinements be properly gauged.

In an attempt to "improve the theory at its basis", the effect of the mutual adjustment of two overlapping bands has been investigated by Band (1946), who obtains a more general form of the inequality ($k\theta'/\epsilon_0 > \frac{2}{3}$) for the occurrence of ferromagnetism, and, with certain approximations, a simple "universal" form for the equation relating θ/θ' and $\epsilon_0/k\theta'$, which embodies the effects of both the s and d bands. With the aid of the universal equation, he is able to draw very interesting tentative conclusions about the band head forms in ferromagnetic metals and alloys. It is, however, difficult to judge whether these conclusions are consequential on assumptions which would be generally acceptable or on others, implicit in the work, which seem to require fuller justification; for this reason, a short account of the work might be misleading, and reference should be made to the original papers.

In a report (1940) presented at a conference on magnetism at Strasbourg in 1939, Néel gives an admirable survey of the experimental work on ferromagnetic metals and alloys, and he discusses various possible theoretical approaches to the problems presented, incorporating, in very summary form, a number of his own investigations on rectangular bands and on what may be called ionic equilibria in metals. He comes to the conclusion that if the magnetization and susceptibility curves of nickel, in particular, are "corrected" for the effects of fluctuations and thermal expansion, the agreement with the standard curves for independent electron spins (with $j = \frac{1}{2}$) is not so close as for coupled spins with $j = 1$ or even $j = \frac{3}{2}$. This might be explained by a coupling of electrons within the d band. The possibility is also explored of treating the metal as a mixture of ions in dynamical equilibrium bathing in an atmosphere of 4s electrons; in nickel the ions would be partly neutral atoms with zero moment, and partly ions in the d^8 configuration, with a moment corresponding to two coupled spins. This rather reactionary point of view, which Néel appears to favour, seems to meet the diverse requirements which emerge from his analysis of the experimental data rather more satisfactorily than a pure band treatment. None the less such a view can hardly be adopted with enthusiasm, not so much because the treatment is "wrong", as because it corresponds to the use of a very much cruder first approximation to the electronic structure of a metal than is provided by the energy band treatment. This is one of the provocative "ébauches de théorie" with which Néel provisionally links together groups of experimental facts, bringing out clearly how incomplete is present understanding of them.

It is perhaps worth while to draw attention to the fact that the theoretical calculation of the electronic properties of a metal involves the solution of two distinct problems. There is first the problem which may be expressed as that of determining the energy distribution of states; and, second, the problem of determining the distribution of electrons among those states, as dependent on temperature and other conditions. (The particular difficulty about ferromagnetism, a cooperative phenomenon, is that the energies associated with the states are dependent on the distribution of electrons among them.) In this section it is the second problem which has been in the forefront, and an indication has been given of how a complete solution has been obtained for a particular representative distribution of states corresponding to a parabolic band form. Although the results are applicable in detail only to single or superimposed bands of this type,

their general character is unlikely to be critically dependent on details of the band form. The general methods used in computing the required Fermi-Dirac functions could be applied to bands of any form, including those obtained numerically in the Wigner-Seitz-Slater type of calculation, and exemplified in figure 3.2. The computation of such band forms is a remarkable achievement, but at present the uncertainties in detail are so great as to make the labour which would be entailed in statistical calculations for these particular band forms completely disproportionate to the value of the results.

The most detailed calculations which have been made on electron energy bands in ferromagnetic metals are probably those of Manning (1943) on body-centred iron and of Greene and Manning (1943) on face-centred iron. Manning discusses the effect of the various approximations involved in the method used and concludes that the general features of the electronic structure of metallic iron as obtained by the method are essentially correct, but that detailed numerical agreement with experiment is not possible without much more elaborate methods of calculation. In exemplification of this lack of detailed agreement it may be mentioned that the computed electronic specific heat at low temperatures is less than half the value found experimentally (see §4(ii)). The character of the computed band in the neighbourhood of the top of the Fermi distribution appears to be much more complex than for nickel in that in this region there are four overlapping partial bands. This suggests that a treatment of the band structure as consisting of two overlapping parabolic bands would be far less likely to give a good approximation to the experimental results for iron than for nickel. It is rather surprising that the behaviour of iron and nickel is so similar in respect of the variation with temperature of the spontaneous magnetization below, and of the susceptibility above the Curie temperature (see §4(ii)).

It should be finally mentioned that when there is an overlap of bands, the characterization of individual bands or "partial" bands as d bands and s bands, and the electrons as d and s electrons, is an oversimplification. The method of calculation is such that bands ordinarily called d bands would be better described as bands in which d functions predominate. The precise bearing of this on interchange interaction effects is not at present clear. Indeed, a method which is at once satisfactory, from a rigorous standpoint, and practicable, of dealing with interchange interaction effects within the framework of the general collective electron scheme has yet to be developed.

3(v). *Magnitude and temperature variation of interchange interaction*

The important question of the dependence of interchange interaction on interatomic distance has been briefly discussed from a theoretical standpoint in §2(vi). Independently of theory, however, it is clear from the experimental facts that ferromagnetism occurs for certain limited ranges of distances between relevant atoms, whether in metallic elements, alloys or compounds. Many compounds of manganese, for example, such as MnAs and MnSb, are ferromagnetic, but the element is not, the separation of the Mn atoms being greater in the compounds. Again, body-centred iron, where the atoms have 8 neighbours at 2·49 A. and 6 at 2·96 A., is ferromagnetic, but not face-centred iron, with 12 neighbours at 2·57 A. It will be apparent that if the collective electron treatment outlined in §3(iv) is

along the right lines, values of the interchange interaction deduced by the application of a simple treatment similar to that of §2(vii) may be greatly in error. None the less, the general indications provided by a systematic treatment of the experimental results, such as that of Néel (1936, 1940) are probably significant, and are certainly of great interest. For elementary substances which follow approximately a Weiss law, $\chi_A = C_A/(T-\theta)$, Néel calculates the ratio $\theta/C_A z$, where z is normally taken as the number of nearest neighbours of an atom. (The original paper must be consulted for details of the choice made in particular cases.) The ratio, a reduced molecular field coefficient denoted by u, is readily shown to be proportional, on a simple theory (cf. §2(vii)), to the energy associated with the interchange interaction between pairs of electrons in neighbouring atoms. A method is developed by which the corresponding quantity for interaction between different atoms can be obtained from results on alloys. By a special, and rather dubious, hypothesis an estimate is also made of the reduced molecular field coefficient in metals with a paramagnetism approximately independent of the temperature. The values of u are plotted not against the interatomic distance, d, but against $d-2r$, where r is the effective radius (i.e. the radius corresponding to the maximum of the charge density distribution), calculated by Slater's method, of the electron shell, the d shell in the cases considered, contributing the magnetic moment. The results are shown in figure 3.5. As a representation of the

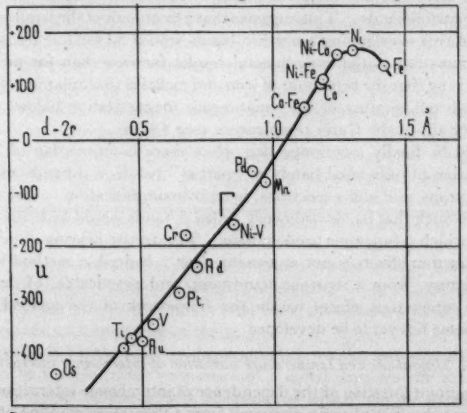


Figure 3.5. Curve representing the variation of the reduced molecular field coefficient, u, with atomic separation. d, interatomic distance; r, effective radius of d electron shell. (Néel, 1940.) (*Reproduced, by kind permission, from* '*Le Magnétisme*'.)

variation of interchange interaction with interatomic distance this curve should be accepted only with many reservations; it is, however, probably the best indication, based on experiment, of the general character of that variation which has so far been obtained. The form of the curve is in agreement with that which had been presented earlier on the basis of a theoretically controlled guess. Néel adduces evidence to show that, although there would presumably be distinct

curves for particular pairs of atoms, these curves are probably similar to the mean curve shown in the diagram. It may be mentioned in passing that Dehlinger (1936–39) has given a series of interchange interaction, distance curves for the individual transition elements on the basis of considerations of the magnetic properties of alloys, in which the possibility is envisaged, in certain cases, of negative interaction between nearest and positive interaction between next nearest neighbours.

The variation of interchange interaction (or of the molecular field coefficient) with interatomic distance for a single substance has been very carefully considered by Néel (1937, 1940), in connection with the possibility of a significant temperature dependence of the molecular field coefficient. He discusses fully various relevant experimental findings—on the "anomalous" temperature variation of the coefficient of expansion in the neighbourhood of the Curie point, the effect of pressure on the Curie temperature, volume magnetostriction in high fields and the like—which it will be impossible to summarize usefully in this report. On many points, owing in part to the smallness of the effects and the lack of precision, or lack of agreement, in the experimental work, no definite conclusions can be reached. Néel shows, however, that a variation of the molecular field coefficient, n, can result from the increase with temperature of the interatomic distance, and of the amplitude of the atomic oscillations. Writing

$$n = n_0(1 + \lambda_1 T + \lambda_2 T^2 \ldots), \qquad \ldots\ldots(3.2)$$

an expression is obtained for λ_1, involving the first and second derivatives of the curve of figure 3.5 and the thermal and elastic constants of the material; the estimated values are $+1{\cdot}1 \times 10^{-4}$ and $-2{\cdot}2 \times 10^{-4}$ for cobalt and nickel respectively. It is then shown that the constants C and θ in the Weiss formula $\chi = C/(T-\theta)$ are related to those, say C_0 and θ_0, in the formula calculated without making allowance for thermal variation, by

$$C_0 = C/(1+\lambda_1\theta); \qquad \theta_0 = \theta/(1+\lambda_1\theta). \qquad \ldots\ldots(3.3)$$

Adopting Néel's estimates of λ_1, the values of $\lambda_1\theta$ for cobalt and nickel are approximately $+0{\cdot}15$ and $-0{\cdot}15$. Similar modifications would be required in the implicit formula for the magnetization below the Curie point. If Néel's estimates of λ_1 are correct, even as to order of magnitude, the effect of temperature variation of the molecular field coefficient (or of the interchange interaction) is far too large to be ignored in attempts to derive effective moments (or numbers of effective electron spins) from inverse susceptibility, temperature curves above the Curie point; but more certain methods of estimating this temperature variation are clearly desirable before the necessary "corrections" can be confidently applied.

In connection with the temperature variation of the molecular field coefficient, reference should perhaps be made to the thermodynamic treatment, developed by Stoner (1935, 36b, 37) which, in principle, allows an estimate of the coefficient to be made over a range of temperatures, not only from the specific heat excess and the magneto-caloric effect, but also from the field and temperature variation of the intrinsic magnetization. This work will not be discussed in detail here, as it was fully reviewed by Bates in an earlier *Progress Report* (1937). A very large apparent variation was revealed (an increase by a factor of 10 in the coefficient for nickel between room temperature and the Curie point), but it was concluded that

this, as well as the major inconsistencies in values obtained by different methods, were mainly due to the change in magnetization even in very high fields (corresponding to the term γH in (2.19)) not being exclusively a change in intrinsic magnetization. If changes in intrinsic and apparent magnetization could be differentiated with greater certainty, and without making assumptions on the point which it is desired to examine, the method of treating the data would seem entirely appropriate for clearing up the present uncertainties. The particular theoretical problem of the variation with field of the intrinsic magnetization has been treated in great detail on the Heisenberg basis by Holstein and Primakoff (1940). They are not primarily concerned with the question under consideration here, but their critical discussion of some of the experimental data is very relevant to it.

§ 4. INTRINSIC MAGNETIZATION. EXPERIMENTAL

4 (i). *Introductory*

Experimental investigations of the quasi-saturation magnetization of a ferromagnetic below the Curie point and of the susceptibility above may be expected, in the light of the theory, to provide information about the elementary carriers of the magnetic moment and the interchange interaction between them. In the ferromagnetic metals, the evidence indicates that the carriers are essentially electron spins with a moment of 1 Bohr magneton. Usually all the spins which contribute to the ferromagnetism will be aligned parallel to each other at absolute zero (though the possibility of incomplete alignment in special cases, as discussed in § 3 (iv), should be kept in mind), and the number per atom, which will be denoted by q, can be derived at once from the spontaneous (or, effectively, the saturation) magnetization at absolute zero; this is obtainable with considerable precision by suitable extrapolation of the observed magnetizations in high fields at low temperatures. Denoting the number in the ferromagnetic state derived from the low temperature saturation magnetization by q_f,

$$q_f = (\sigma_A)_0/M_B = A\sigma_0/M_B = AI_0/\rho M_B, \qquad \ldots\ldots (4.1)$$

where A is the atomic weight, ρ the density, M_B the gram molecular Bohr magneton. ($M_B = 5{\cdot}585 \times 10^3$; $1/M_B = 1{\cdot}790_4 \times 10^{-4}$.) The modifications required in (4.1) for alloys are obvious.

In the region above the Curie point the variation of the susceptibility with temperature may be represented empirically, sometimes over considerable ranges, by a Curie-Weiss law,

$$\chi_A = C_A/(T-\theta), \qquad \ldots\ldots (4.2)$$

and the results may be summarized in the values of C_A and θ appropriate for the particular temperature range. Owing to the marked curvature, concave upwards, of the $1/\chi$, T graphs invariably found over some tens of degrees in the neighbourhood of the Curie point, and the less pronounced curvatures (sometimes concave, sometimes convex) usually revealed at higher temperatures when an extended temperature range can be examined, the behaviour of ferromagnetics above the Curie point cannot ordinarily be completely specified by unique values of C_A and θ. It is customary to denote the value of θ obtained by extrapolation of an approximately linear portion of a $1/\chi$, T curve by θ_p, and to use θ_f for the temperature at which spontaneous magnetization is presumed to disappear. Provided

that the points mentioned above are kept in mind, the summarizing of experimental results in the values of C_A and θ_p is unobjectionable, and it may be pushed somewhat further. If Maxwell-Boltzmann statistics (in the "quantized" form) rather than Fermi-Dirac were applicable, and if, further, fluctuation effects were negligible, a linear relation between $1/\chi$ and T would be expected theoretically and the number of electron spins per atom could be calculated from C_A. If there are q_1 independent spins per atom with moment μ_B, q_2 coupled in pairs with moment $2\mu_B$ for each pair, and generally q_r with moment $r\mu_B$ for each r, it may be shown that, for the conditions assumed,

$$\sum_{r=1} q_r(r+2)/3 = RC_A/M_B^2. \qquad \ldots\ldots(4.3)$$

Denoting the number of effective spins per atom calculated from the paramagnetic susceptibility on the assumption that each carrier consists of r spins in parallel coupling by $(q_p)_r$,

$$\tfrac{1}{3}(r+2)(q_p)_r = (q_p)_1 = RC_A/M_B^2 = 2{\cdot}665C_A. \qquad \ldots\ldots(4{\cdot}4)$$

Where electron spins only are concerned, the specification of $(q_p)_r$ seems more appropriate, as giving an immediate indication of the effective number of spins, than the usual Bohr magneton number p_B defined by

$$p_B^2 = 3RC_A/M_B^2. \qquad \ldots\ldots(4.4\,a)$$

The number of parallel spins per atom derivable from the low temperature saturation magnetization by (4.1) is independent of whether or not there is coupling of the spins. If the conditions obtained for the approximate applicability of the simple modified molecular field treatment of §2(vii), with the results illustrated in figure 2.4, for which independent spins with $j = \frac{1}{2}$ are assumed, the values of q_r from (4.1) and $(q_p)_1$ from (4.4) would agree. Under similar conditions, but with carriers consisting of r spins coupled in parallel, $(q_p)_r$ would be the same as q_r. If appreciable changes in the atomic arrangement and electronic structure of the material in passing from low temperatures to temperatures above the Curie point can be excluded, the lack of agreement usually found can properly be attributed to inadequacy of the simple theory, and must be considered in the light of the later theoretical developments outlined in §3. For purposes of a general survey of experimental results an equation such as (4.4) is extremely useful in giving immediately an approximate indication of the atomic significance of the measured susceptibilities.

The Curie temperature itself gives an indication of the magnitude of the interchange interaction, but no finality has yet been reached as to just how $k\theta$ is related to the interchange interaction between a pair of electrons. This is partly owing to the uncertainty as to which and how many neighbours are to be regarded as in effective interaction, and partly owing to the very definite indication of the collective electron treatment that positive interchange interaction does not necessarily result in ferromagnetism and a positive Curie temperature. Although from this particular standpoint the details of the behaviour of a ferromagnetic in the Curie point region may be disregarded, from other standpoints, both theoretical and experimental, they are of great interest and importance. The question is whether there is, in fact, an unambiguously definable Curie point at which the spontaneous magnetization disappears completely and abruptly

(as is suggested by a curve on the scale shown in figure 2.1). The general experimental indications are that there is a real "tailing", as it always appears in the direct experimental observations, whether of magnetic quantities (such as magnetization in high fields or initial susceptibility) or of associated quantities (such as specific heat, or the temperature coefficient of the electrical resistance). The whole question has been very fully discussed by Gerlach (1939), who concludes that, when proper precautions are taken, values which are obtainable for a Curie temperature defined as the temperature at which diverse physical properties show a maximum "anomaly" (connected with the maximum value of $\delta\sigma/\delta T$ where σ is the spontaneous magnetization) are for nickel consistent to within 2°, but that above this temperature anomalies presumably associated with a small decreasing localized spontaneous magnetization persist for some 50°. With alloys the tailing even in curves of the type of those in figure 2.1 is much more pronounced, and the precision possible in the determination of a Curie point by any procedure is correspondingly reduced. There would be general agreement about tailing in alloys, which is connected with the great difficulty, even impossibility, of ensuring strict homogeneity of composition. Whether for ideally pure and homogeneous metals the tail would completely disappear is perhaps an academic question. A full list of papers on this topic is given by Gerlach. Among the more recent papers reference may be made particularly to those of Néel (1935), giving detailed measurements on the susceptibility of nickel in the immediate vicinity of the Curie point, and of Kussmann and Schulze (1937) and Potter (1937), dealing with the temperature variation of resistance in the Curie point region.

The above discussion correctly implies that there is considerable uncertainty about the precise significance of "Curie points" quoted in the literature, an uncertainty quite distinct from those arising from experimental errors and impurity of the specimens. The uncertainty is often of no particular consequence, but where it is, it can be removed only by a sometimes very troublesome search in the literature for an account of the relevant original work. In the absence of information to the contrary it may be assumed as most probable that the Curie temperature has been derived by extrapolation to zero magnetization of the steepest part of a (σ, T) or a (σ^2, T) curve obtained with relatively low fields. Measurements of the magneto-caloric effect in the neighbourhood of the Curie point make possible the derivation of a spontaneous magnetization, temperature ($(\sigma_T)_0$, T) curve, and the temperature obtained by extrapolation of the steepest part of the $(\sigma_T)_0^2$, T curve has perhaps a better claim to be regarded as *the* Curie temperature than any other. But this curve, too, has a tail, and elaborate magneto-caloric investigations such as those of Weiss and Forrer (1926) on nickel, and of Potter (1934) on iron, obviously could not be undertaken merely for the determination of a Curie temperature.

In recent years a considerable amount of experimental work has been carried out on the variation with composition of the saturation moment and Curie temperature of ferromagnetic alloys and of the Curie constant of both ferro- and paramagnetic alloys, and results of great interest have been obtained. It may perhaps be appropriate to mention here explicitly in connection with such work that not only the composition but also the phase constitution of the

alloys must be known if significant conclusions are to be drawn from the results. What is first required is a knowledge of how the magnetic characteristics vary with composition and temperature over single phase regions, e.g. how the saturation moment per atom (or the q_t value) varies over the range of solid solution of element B in element A. Once knowledge of this kind is obtained, whether or not it receives a theoretically satisfactory interpretation, it provides a basis for the application of magnetic methods to the study of structural changes and other transformations in alloys, methods which can sometimes give information which is not readily obtainable by any of the standard metallurgical or x-ray methods. Unfortunately, only the barest indication of the use of such methods can be given under (v) below. The more direct magnetic studies of alloys, mainly binary, will be dealt with under (iii) and (iv), and the ferromagnetic elements under (ii).

4(ii). *The ferromagnetic elements*

Magnetization and magneto-caloric effect. There seem to have been no more recent absolute measurements aiming at the determination of the low temperature saturation moments of Fe, Co and Ni comparable in precision with those of Weiss and Forrer in 1929, and the values obtained in their work are given in table 4.1 below. Their value for the specific saturation magnetization of pure iron at room temperature (288° K.) is practically the same as that obtained by Steinhaus, Kussmann and Schoen (1937)—$217{\cdot}7_6$ as compared with $217{\cdot}9_8$. The detailed study by the same investigators (1926) of the magnetization and of the magneto-caloric effect for Ni has also not been superseded. A comparable study of Fe has, however, been made by Potter (1934), who developed ingenious new experimental methods having several advantages over those previously used. He discovered the apparent variation with temperature of the molecular field coefficient as derived from the magneto-caloric results to which reference has been made in § 3 (v). Potter also constructed from his results a corresponding state curve (σ/σ_0 against T/θ) for Fe. This is closely similar to, but not identical with, the curve for Ni, with rather larger deviations over the range $0{\cdot}7 < T/\theta < 1$ from the curve given by the simple theory for independent electron spins (figure 2.4).

Specific heat. Many investigations on the specific heat of Ni have been made, the most recent, and probably the most accurate, being that of Sykes and Wilkinson (1938). Measurements were made on four types of nickel, including two of better than 99·9 % purity, over a temperature range from 60° to 600° C. the Curie point region being studied in detail. The effects of differences in composition, method of manufacture, and heat treatment were examined. The results, together with those of previous investigators, were critically considered, and a "most probable" specific heat curve evaluated. The results show that the maximum specific heat at the Curie point is approximately 3·0 cal. deg^{-1} mole^{-1} in excess of the lattice vibration value, and that the persistent excess, in the main almost certainly an electronic specific heat, above the Curie point is about 1·2. Although the decrease from the maximum is at first very rapid (by more than 1 cal. deg^{-1} mole^{-1} in 2°), it seems to be quite definitely established that the tail extends over some 30°. The results were analysed by

the method used by Stoner (1936 b, c) for earlier results; the method is generally suitable, though modifications in detail are necessary in the light of later theoretical developments.

The most recent extensive investigation of the specific heat of Fe is that of Awbery and Griffiths (1940), on a very pure specimen (better than 99·99%) for temperatures extending from 50° to 950° c., the thermal capacity being measured over ranges of 5° c. or less. The "magnetic" maximum occurred at 755° c. Above this temperature there is a rapid falling off, but the thermal effect of the magnetic transformation (what has above been called the "tail") is still visible when the β-γ transformation, a latent heat change giving a very high narrow peak in the apparent specific heat curve, sets in at about 903° c. A specific heat curve for iron covering the Curie point region, and a curve for nickel are shown in figure 4.1, drawn from the data of Awbery and Griffiths and of Sykes and Wilkinson respectively.

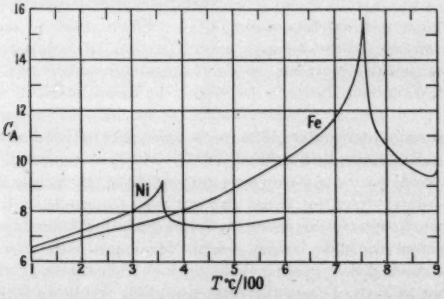


Figure 4.1. Specific heat of nickel (data from Sykes and Wilkinson, 1938) and iron (Awbery and Griffiths, 1940). C_A, gram atomic heat, cal. deg^{-1} mole^{-1}.

Measurements of the specific heat at very low temperatures, usually over the range from about 1° to 20°K. have been made on Ni (Keesom and Clark, 1935), Co (Duyckaerts, 1939) and Fe (Keesom and Kurrelmeyer, 1939). At these temperatures the specific heat is closely represented by

$$C=\gamma T+bT^3, \qquad \ldots\ldots(4.5)$$

where the second term is that associated with lattice vibrations (the Debye term), and the first corresponds to an electronic specific heat. When C is expressed in cal. deg.$^{-1}$ mole^{-1} the values for $\gamma\times10^4$ are: Ni, 17·44; Co, 12·0; Fe, 12·0. These are much higher than the values found for "normal" metals (e.g. for Ag the value is 1·6, in close agreement with the value calculated theoretically for "free" electrons, one per atom) and indicate a correspondingly higher energy density of states at the top of the Fermi distribution. The higher values fit in well with the general energy band treatment of the transition metals (Mott and Jones, 1936, 37). The relation between the electronic specific heat at low and high temperatures has been treated quantitatively by Stoner (1938 b, 1939 a) for the standard band form (see § 3 (iv)), and leads to a value for the excess specific

heat of nickel above the Curie point in reasonably good agreement with that deduced from the experimental results. It should be said, however, that anything beyond a superficial examination of the specific heat data for ferromagnetics, whether at low temperatures or over the whole range, soon reveals that little more than a promising beginning has yet been made in the quantitative theoretical interpretation of the results.

Paramagnetism. Two comprehensive experimental investigations, both models of their kind, have recently been made on the susceptibility of Fe, Co and Ni above the Curie point by Sucksmith and Pearce (1938) and by Fallot (1944). Different types of balance were used, a suitably adapted form of his ring balance by Sucksmith, and the Weiss-Foëx-Forrer translational balance by Fallot (who completed the bulk of his measurements at Strasbourg before 1940). In view of the many difficulties in detail in magnetic measurements at high temperatures (up to 1500° c.), the fact that different methods were used makes even more satisfactory the very close agreement in the results obtained. The results, as far as they can be usefully represented in a small scale diagram, are shown in figure 4.2. (A more extended curve for nickel is shown in figure 2·2.)

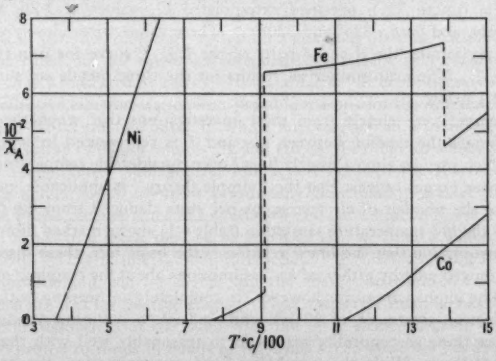


Figure 4.2. Susceptibility of iron, cobalt and nickel above the Curie point. χ_A, gram atomic susceptibility. (*From the data of Sucksmith and Pearce*, 1938, *and Fallot*, 1944.)

In the $1/\chi$, T representation a curvature, concave upwards, occurs for all three metals in the neighbourhood of the Curie point, the region examined in detail for Fe and Ni by Néel (1932, 35), and this is followed by an appreciably linear portion extending over nearly 400° for Ni, and 200° for Co. The values of C_A obtained from these regions in the two investigations agreed to within 1%. For Co, Sucksmith found a marked dependence of susceptibility on previous heat treatment, associated with density variations; the θ_p values for different specimens ranged from 1130 to 1155° c. (Fallot's value is 1156° c.), but the C_A values were the same. (It may be noted that Co is face-centred cubic at these temperatures, and at room temperature normally close packed hexagonal. An x-ray study of the transformation, which is of a peculiar type, has been made

by Edwards and Lipson, 1943.) At higher temperatures the $1/\chi$, T curve for Ni shows a definite curvature concave downwards (see figure 2.2), but there is no break in the continuity of the curve as a whole; Fallot, in fact, refers to the line from which the lower temperature C_A value is obtained as an inflexional tangent. Iron is body-centred cubic below about 900° c. (the β phase) and above 1390° c. (the so-called δ phase), and face-centred (γ) between. The β–γ and γ–δ transformations are of the ordinary type, taking some hours for completion, and showing temperature hysteresis. The much smaller susceptibility in the γ state is obvious from figure 4.2. The $1/\chi$, T curve is nearly linear (there is a slight concavity upwards) but no direct significance can be attached to the values of C_A (about 6·7) and θ_p (about $-3000°$ K.) which can be derived. Owing to the limited range of the β phase the derivation of C_A and θ_p values is somewhat hazardous. The β–γ transition temperature can, however, be raised, or the γ phase region can be entirely eliminated, by alloying iron with certain soluble non-magnetic metals. A longer nearly linear $1/\chi$, T range is then obtained, from which C_A and θ_p values can be derived, and by extrapolating to zero content of the added metals, reasonably precise estimates of C_A and θ_p may be made for pure iron. Sucksmith, using alloys with vanadium (up to 5 atomic per cent), and Fallot, using alloys with aluminium (up to 7%), obtained extrapolated C_A values of $1{\cdot}26_5$ and $1{\cdot}26_3$ respectively, and for θ_p, 828° and 820° c. These values agree closely with those obtained by joining the β and δ parts of the $1/\chi$, T curve for iron (C_A, 1·23; θ_p, 820° c.). The main numerical results for the three metals are summarized in table 4.1 below.

It emerges very clearly from these investigations that when there are no phase changes the relation between $1/\chi$ and T is represented by a continuous curve, which may be almost exactly linear over considerable temperature ranges. These linear ranges suggest that the "simple theory" is applicable, but a comparison of the number of electron spins per atom deduced from the C_A values and from the low temperature saturation (table 4.1) shows marked discrepancies. On the assumption that the linear relation is the basic fact, these discrepancies can be removed only by rather *ad hoc* assumptions about the coupling of electron spins. The slight curvature, however, is completely in keeping with the collective electron treatment of § 3 (iv), and, indeed, a preliminary analysis has shown that these susceptibility results fit in reasonably well with that scheme (Stoner, 1938 c). With the parabolic band approximation the indications are that the number of holes per atom in the d band is slightly in excess of the q_f value. None of the theories accounts immediately for a converse curvature such as is found for Ni at the higher temperatures (see figure 2.2), but it has been shown by Fallot and others that this can be "removed" by allowing for the presence of a constant paramagnetism of suitably chosen magnitude in addition to the main varying paramagnetism. From a collective electron standpoint an effect of the required kind could arise from the s electrons, and also from the effect of a temperature variation of the number of "holes" in the d band (equal, in nickel, to the number of s electrons), but preliminary calculations indicate that the additional paramagnetism arising in this way, though of the right order, is probably too small to account entirely for the observed results.

Gadolinium. In 1935 the exciting discovery was made by Urbain, Weiss

and Trombe that the metal gadolinium (atomic number, 64) is ferromagnetic. Not merely is it ferromagnetic, but at low temperatures the specific magnetization is considerably higher than that of iron. A fairly complete study of the thermomagnetic behaviour has been made by Trombe (1937) using the general methods and procedure developed at Strasbourg. The ferromagnetic Curie temperature, θ_f, is obtained by extrapolation of a (σ^2, T) curve as $(16 \pm 2)°$ C. Below the Curie point measurements of the magnetization were made in fields up to 19 900 oersted at temperatures down to 77° K. The saturation magnetization at absolute zero, found by the usual methods of extrapolation, corresponds very closely to $7\mu_B$ per atom ($q_f = 7{\cdot}1_4$). Above the Curie point the $1/\chi$, T graph shows the usual initial curvature, followed by an appreciably linear portion extending from 90° to 366° C. The Curie constant C_A corresponds very closely to a carrier with a resultant spin moment of $7\mu_B((q_p)_7 = 6{\cdot}9_7)$.

Gadolinium is much simpler than the other ferromagnetic metals in that the saturation moment and the susceptibility variation can both be accounted for by assuming the carrier to be the same as that in the extensively studied salts, that is, the Gd^{3} ion with 61 electrons. The spectroscopic state of the ion is $^8S_{7/2}$, corresponding to a coupling of the 7f electrons of the half-filled f group, the resultant orbital moment being zero, and the total resultant moment $\frac{7}{2}$, due to 7 spins in parallel coupling. Although a coupling of spins within the atoms cannot be ruled out for the other ferromagnetic metals, it is certainly not so unambiguously indicated by the experimental facts. It is of interest to consider, from a collective electron standpoint, possible reasons for the difference between Fe, Co and Ni on the one hand, in which ferromagnetism is associated with an incomplete group of d electrons ($l = 2$), and Gd, on the other, in which the f group ($l = 3$) is incomplete. The general interaction between the relevant electron groups in neighbouring atoms is presumably considerably greater in the first case than in the second, since the overlap of the wave functions will be greater. Now the bands in a metal may be regarded as arising from a widening of the atomic levels owing to inter-atomic interaction. If this interaction is relatively small, the widening will be slight, and the energy distribution of states will approximate to that for the free atoms. Thus in the rare earth metals a widening of the levels associated in the atom with the three electrons in 5d or 6s states may be expected to result in bands of the usual type, while the bands associated with the 4f electron states may be sufficiently narrow for the intra-atomic interaction to predominate over the inter-atomic. In the ferromagnetic metals of the first transition series, the calculations indicate (cf. figure 3.2) that the d band, though narrow compared with the s band, is none the less of considerable width. The question of whether there is coupling of spins within an atom in a metal (at least when there is more than one effective electron per atom) is a question of the relative magnitudes of the intra- and inter-atomic interactions. For the 4f group in Gd the first effect is predominant. For the 3d group in the other ferromagnetics the two effects may be comparable, and this might well account for the failure of any relatively simple theoretical treatment to cover the results in quantitative detail.

In the course of a comprehensive investigation Klemm and Bommer (1937) obtained all the rare earth metals, with the exception of those of atomic number

61 (Il) and 67 (Ho), by reduction of the chlorides, and made measurements of the magnetic susceptibility at temperatures ranging from 90° K. to about 550° K. References to previous measurements on rare earth metals are given in the paper. Some of these metals are strongly paramagnetic, and follow approximately a Curie-Weiss law with a positive θ_p over a wide range. For Tb, 68, and Dy, 69, for example, an approximately linear variation of $1/\chi$ with T over some 300° gives θ_p values of 205° K. and 150° K. respectively. In no case, however, except for Gd, were there conclusive indications of ferromagnetism above 90° K. Where the θ_p value is higher than this, there is a fairly abrupt change in the slope of the $1/\chi$, T curve, and in the lower temperature range the susceptibility varies relatively slowly with temperature. A dependence of susceptibility on field is, however, observed in such ranges. The occurrence of some degree of spontaneous magnetization cannot, therefore, be ruled out even above 90° K., and it may set in below that temperature. While Gd is the only rare earth metal for which ferromagnetism is definitely established, and certainly the only one for which the ferromagnetic Curie temperature, θ_f, is in the ordinary temperature range, until more detailed measurements are made, particularly in the lowest temperature ranges, the possibility of ferromagnetism for some of the other rare earth elements cannot be entirely excluded.

Table of characteristics. It may be useful to give a tabular summary of the main characteristics of the four ferromagnetic elements. For Fe, Co and Ni the data for saturation magnetization are from Weiss and Forrer (1929), the paramagnetic data from Sucksmith and Pearce (1938), the ferromagnetic Curie temperature for Ni from Weiss and Forrer (1926), and the more uncertain θ_f values for Fe and Co from Fallot (1936, 1944). All the data for Gd are from Trombe (1937). The specification of the number of effective spins by q_f and $(q_p)_r$ is discussed in connection with equations (4.1) and (4.4).

Table 4.1. Characteristics of the ferromagnetic elements

	Fe	Co	Ni	Gd
Z	26	27	28	64
A	55·84	58·94	58·69	157·3
Ferromagnetic				
θ_f °c.	770	1120	358	16
σ_0	221·7	162	57·6	254
q_f	$2{\cdot}21_7$	1·71	0·605	$7{\cdot}1_4$
Paramagnetic				
T range °c.	—	1230 1450	500 850	90 366
θ_p °c.	820	1155	377	30
C_A	$1{\cdot}26_5$	$1{\cdot}22_8$	$0{\cdot}32_2$	$7{\cdot}8_4$
$\frac{1}{3}(r+2)(q_p)_r$	3·37	3·27	0·86	20·9

Z, atomic number;
A, atomic weight;
θ_f, ferromagnetic Currie temperature;
σ_0, specific saturation magnetization extrapolated to absolute zero;
q_f, corresponding number of parallel electron spins;
T range, range over which the relation $C_A/(T-\theta_p)$ holds approximately;
θ_p, paramagnetic Curie temperature;
C_A, Curie constant per gram atom;
$(q_p)_r$, corresponding number of electron spins, coupled to give a carrier moment of $r\mu_B$. (For Gd, if $r=7$, $q_p=6{\cdot}9_7$.)

4(iii). *Saturation magnetization of ferromagnetic alloys*

General. Extensive measurements have been made during recent years, particularly at Strasbourg, on the saturation magnetization and the Curie temperature of binary alloys which are solid solutions of a non-ferromagnetic in a ferromagnetic metal. The first comprehensive investigation of this kind was that of Sadron (1932), though previously a number of particular alloys had been studied, and the inter-ferromagnetic alloys had been examined in considerable detail. Sadron's work was mainly on nickel alloys, for which he found in many cases a linear variation of the saturation moment per atom with composition over a considerable range, the moment usually decreasing with increase in content of the non-ferromagnetic element. It has been found most convenient to express these, and other results, by giving the initial slope of the curve relating mean atomic moment, in Bohr units, and atomic composition; that is, with the symbolism used above, by (dq_f/dc) where c is the atomic concentration of the second element. For copper-nickel alloys (investigated by Alder in 1916) the saturation moment decreases almost linearly from the nickel value ($q_f = 0{\cdot}6$) to zero at about 60 atomic per cent of copper, and (dq_f/dc) is $-1{\cdot}0$; each substituted Cu atom reduces the saturation moment by 1 Bohr magneton. This, though an almost misleadingly simple case, illustrates the immediate significance of the numbers obtained when the results are expressed in this form. In table 4.2, the initial values of (dq_f/dc) are given for most of the alloys so far examined from this point of view. The values have been obtained from the data of the original papers, but except for a few additions, are substantially the same as those given in a list by Néel (1940). Where a reasonably definite value for the change of the Curie temperature could be obtained from the recent data, this has been added. The sources of the data are as follows:

Inter-ferromagnetic alloys: Fe–Co, Co–Ni, Weiss and Forrer (1929); Fe–Ni, Peschard (1925).

Alloys with non-ferromagnetic as solute in ferromagnetic solvent:

Fe alloys: Fallot (1936–38), except (dq_f/dc) for Mn, Sadron (1932).

Co alloys: Al, Cr, Mo, W, Farcas (1937); Si, Mn, Pt, Au, Sadron (1932).

Ni alloys: Marian (1937), except (dq_f/dc) for Cu, Alder (1916); V, Mn, Ru, Pd, Sn, Sadron (1932); Sb, Ta, Rado and Kaufmann (1941).

Although the measurements from which the numbers in table 4.2 are derived are in general of considerable accuracy, there may be appreciable errors in the numbers themselves as representing the initial variation of q_f and θ_f over the solid solution range. The starting point is the pure metal, the crystal structure of iron, nickel and cobalt being body centred cubic, face-centred cubic, and hexagonal close-packed respectively. Where there is no change of structure over a wide range, and the variation is nearly linear, the values of (dq_f/dc) can be obtained with fair precision, but where the range of solid solution is narrow, even slight errors in the analyses of the specimens and inhomogeneity effects may give rise to a scattering of the points, which makes an accurate determination of the initial slopes impossible. For some alloys the saturation magnetization is markedly dependent on heat treatment, the significance of which, in terms of

Table 4.2. Change of saturation magnetization and Curie temperature in alloys with ferromagnetic elements.

		Solvent q_f θ_f °K	Fe 2·21$_7$ 1043		Co 1·71 1393	Ni 0·60$_5$ 631	
Solute	Z	Col.	$-\frac{dq_f}{dc}$	$-\frac{1}{100}\frac{d\theta_f}{dc}$	$-\frac{dq_f}{dc}$	$-\frac{dq_f}{dc}$	$-\frac{1}{100}\frac{d\theta_f}{dc}$
Al	13	3	2·3	1·6	2·4	3	27
Si	14	4	2·2	3·5	5·3	3·4	34
Ti	22	4				3·8	33
V	23	5	2·2	−7·5		5·2	56
Cr	24	6	2·2	1·5	7	4·4	42
Mn	25	7	2·2	15	4·7	−2·4	
Fe	26	8			−1·6	−2·8	
Co	27	9	−1·2			−1·2	
Ni	28	10	−0·6		0·9		
Cu	29	1				1	10
Zn	30	2	2	6		2·1	21
Mo	42	6			6	6	
Ru	44	8	0	16		2·8	29
Rh	45	9	−1	2			
Pd	46	10	0·2	3		0	
Sn	50	4	2·2	0		4·2	
Sb	51	5				3·6	
Ta	73	5				5·7	
W	74	6			6	6	
Os	76	8	1·6	11			
Ir	77	9	−0·9	4			
Pt	78	10	−2	0	1·1	0·6	
Au	79	1	2·2	0	1·2	0·8	12

Z, atomic number;
Col., column of periodic table;
q_f, saturation moment per atom in Bohr magnetons (number of parallel electron spins per atom);
θ_f, ferromagnetic Curie temperature;
c, atomic concentration of solute metal.

phase structure changes, could be determined only by extensive special investigations (see §4(v)). In a few cases possible changes of structure with composition were not considered in detail; the values for the Co alloys, in particular, must be regarded as preliminary, and as not necessarily having reference purely to the hexagonal structure range. It is not possible to enter into details about each alloy here, but it may be said that for most of the iron and nickel alloys the values definitely correspond to a single phase region; the uncertainty in the (dq_f/dc) values is of the order 0·1, and in the $(d\theta_f/dc)$ values from 5 to 10 %.

Some of the alloys have been studied in much greater detail than may be suggested by the table; an excellent recent example is provided by the work of Fallot (1938) on the alloys of iron with the metals of the platinum family.

Nickel alloys. For the nickel alloys with elements on the left hand side of the periodic table, up to columns 5 or 6, a very simple relation is at once apparent from table 4.2. The reduction in moment (in Bohr magnetons) per substituted atom is approximately equal to the number of relatively loosely bound electrons, or effective valency number, of the atom. A plausible interpretation on the basis of the collective electron treatment was first given in a convincing form by Mott (1935) (see § 3 (iv)). It is assumed that the electronic equilibrium in the alloys is such that the number of electrons per atom contributing to the s band remains at first approximately constant and equal to the number in nickel, about 0·6. The more loosely bound electrons above this number in the substituted atom fill up the holes in the d band, the direction of the spins necessarily being such that the resultant moment is correspondingly reduced. Of the v electrons contributed by the atom to the s and d bands, 0·6 enter the s band and $(v-0{\cdot}6)$ the d band, in place of the 0·6 electrons and 0·6 holes contributed by the nickel atom. The moment is accordingly changed by $0{\cdot}6-(v-0{\cdot}6)$, that is, by $-v$; the observed values of (dq_t/dc) show that v may in many cases be identified with the column number of the element. It may be noticed that since the height, or density of states, of the d band in nickel at the top of the Fermi distribution is much greater than that of the s band (see figure 3.2), there is some theoretical as well as empirical justification for supposing that the number of s electrons per atom will not change appreciably until the d band is full. Even so, it is remarkable that so simple a scheme covers the observations as closely as it does for the nickel alloys with elements of columns 1 to 5 of the periodic table. Attempts at a more detailed theoretical treatment of the nickel alloys, based on Slater's method of estimating the interchange interaction effects (see § 3 (iv)), have been made by Hirone (1938), and developed in some directions by Niessen (1939). Niessen considers the question of whether, in the alloys, the holes in the d band are to be regarded as associated equally with all the atoms or with the nickel atoms only. The change in mean moment does not differentiate between the two possibilities. Niessen's argument suggests that the variation of the Curie temperature may provide a differentiating criterion. It may be seen from table 4.2 that the Curie temperature would in some cases extrapolate linearly to zero at roughly the same composition as the moment. In such cases, particularly for the Cu and Zn alloys, the argument would indicate a common d band, with a sharing of the holes among all the atoms, but in other cases a closer association of the holes with the nickel atoms. For alloys, the theoretical problem of the energy and space distribution of those electrons which cannot be regarded, even approximately, as associated with particular atoms, is much more difficult than for metals owing to the differences in the charge distributions and fields round the two or more kinds of atoms present. Until this general problem is resolved more completely than it is at present any quasi-theoretical consideration of the ferromagnetic alloys must necessarily be of a tentative character.

Iron alloys. For the iron alloys, the simple tentative scheme which is

reasonably satisfactory for the nickel alloys does not work at all. For all the elements investigated except those of columns 8, 9 and 10, the reduction in q_f per substituted atom is approximately the same as the q_f value for iron, 2·2. The initial variation is as if atoms of moment 2·2 were replaced by atoms of zero moment. From a collective electron standpoint a natural interpretation of the difference between iron and nickel is that in Fe saturation at absolute zero is incomplete, so that the number of holes in the d band is greater than the q_f value. Not only has this possibility been envisaged, but the necessary conditions have been fully treated quantitatively for a band of parabolic form (see § 3(iv)). It is reasonable to suppose that the character of these conditions is not peculiar to the special band form studied in detail, and that in general there will be a lower limiting value for the interchange interaction for the occurrence of any degree of spontaneous magnetization, and a higher lower limit for the complete alignment of the available electron spins. In iron, taking the q_f value as 2·2, it may be supposed that the number of holes per atom in the d band is $2{\cdot}2+r$, and the number of s electrons $0{\cdot}2+r$. (If q is the number of holes in the d band, then $q_f=q\zeta_0$, and $r=q-q_f=q_f(1-\zeta_0)/\zeta_0$.) An obvious adaptation of the treatment given for nickel then indicates that there will be no change in the moment per Fe atom until the r holes in the d band are filled by electrons in pairs with oppositely directed spins, and that $-(dq_f/dc)$ will have an initial value equal to q_f for iron, that is 2·2. No very definite estimate of the value of r has as yet been made. A preliminary examination of the alloy data for the range of composition over which the initial rate of decrease of moment is maintained indicates that r is not less than 0·3. The tentative conclusion from the alloy results that the saturation in iron is incomplete receives some support from the variation of susceptibility above the Curie point (see § 4(ii)), but perhaps the most convincing qualitative evidence comes from a consideration of the cohesive energy, which indicates that it is very unlikely that the number of s electrons per atom should be as low as 0·2 (see Seitz, 1940, p. 430). A larger number necessarily entails that the number of holes in the d band should be correspondingly greater than 2·2. As against the view, here suggested, it should be stated that the energy band calculations of Manning (1943) for body-centred iron (see § 3 (iv)) give the number of electrons per atom in the s-like band as about 0·25.

The Néel diagram. A very striking graphical representation of the magnetization results for a large number of alloy systems has been given by Néel (1940). His diagram is reproduced in figure 4.3. He assumes for the purpose of the diagram that the measured q_f values in alloys correspond to the number of holes in the d band, and further, that the holes are associated with the ferromagnetic atoms only; he calculates the mean number of holes per ferromagnetic atom, and plots against this the mean number of s electrons per atom, the s electrons being regarded as equally "shared" among all the atoms. It is not possible to discuss here the wealth of detailed information collected together in the Néel diagram. The general points to be noted are the regularity and comparatively narrow range of s values for D values up to about 1·5, and the subsequent irregular changes and the abrupt drop to the s value of 0·2 at iron. The character of the curves suggests that the assumption on which they are drawn, in effect that saturation is complete at absolute zero, is valid in the first region

but not in the second. As Néel points out, the trend is towards an s value of about 0·8 on the iron line; if this is actually the number of s electrons per atom in iron, the q_f value of 2·2 would mean that saturation is incomplete and that of the 2·8 holes in the d band, only 2·2 are parallel at absolute zero. (In terms of electrons iron has 8 in the d and s bands together. With complete saturation there are 5 electrons with positive spin in the d band, and 2·8 with negative, and consequently 0·2 in the s band. The alternative suggested for illustrative

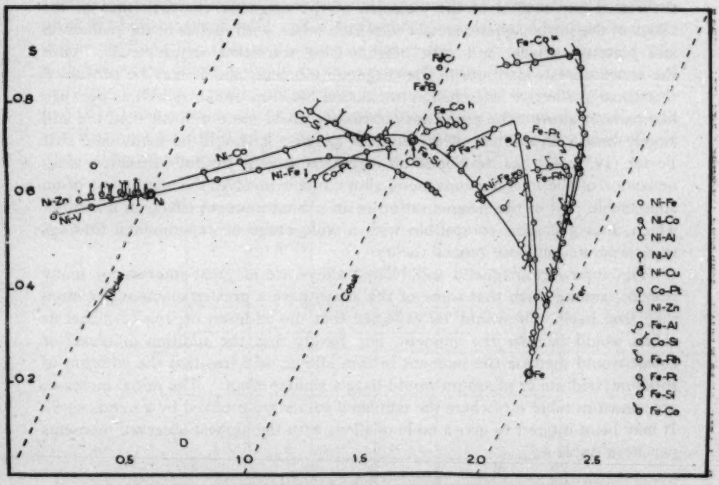


Figure 4.3. Number of s electrons per atom as a function of the mean number, *D*, per ferromagnetic atom, of holes in the d band for ferromagnetic metals and alloys. (Néel, 1940.) (*Reproduced, by kind permission, from 'Le Magnétisme'.*)

purposes is 4·7 positive and 2·5 negative spins in the d band and 0·8 electrons in the s band.) It is not suggested that all the irregularities and abrupt changes of slope in the curves in the Néel diagram could or should be smoothed out by suitable assumptions about the degree of incompleteness of saturation. Phase changes will in general be associated with equally abrupt changes in the character of the electronic equilibrium. Moreover, the interpretive scheme, as presented, is necessarily very limited in scope, and, indeed, superficial, for it does not involve the explicit consideration of how the interchange interaction itself may be modified with changing composition even over a single phase range.

Inter-ferromagnetic and related alloys. The characteristics of the alloys so far considered, that is, of ferromagnetic elements with elements of the left-hand columns of the periodic table, can be coordinated to some extent by the scheme which has been presented. In this scheme the substituted atoms are regarded in a sense as electron donors, providing electrons which fill up the holes in the d band associated with the ferromagnetic metal, and so reduce the moment

per atom. The scheme, it must be understood, is only semi-theoretical ; theoretical knowledge of electron energy band structure in alloys is far too meagre to provide a basis, at present, for general theoretical predictions (even after the event) of how the saturation moment and the Curie temperature "should" vary in alloys. Theoretically it is perhaps more surprising that a simple algebraical scheme can be applied successfully to so many alloys than that it cannot be applied to all of them. There is, at present, no satisfactory quantitative theoretical explanation of the variation of the magnetic characteristics in the alloys of the ferromagnetic metals with each other, with metals of the palladium and platinum triads, and with neighbouring transition series metals. Both the constituents contribute to the magnetic moment, and it may be presumed that there is effective interchange interaction between unlike as well as between like pairs of atoms; the problem is obviously much more difficult than the still largely unsolved problem of the simpler alloys. It should be mentioned that Forrer (1939, 40) has developed an ingenious scheme for the saturation magnetization of the inter-ferromagnetic alloys, but it involves the attribution of an appreciable part of the magnetization to an orbital moment effect, in a manner which does not seem compatible with a wide range of experimental findings as incorporated in more general theory.

The inter-ferromagnetic and related alloys are of great interest for many reasons, among them that some of the alloys have a greater moment per atom than iron itself. It would be expected that the addition of iron to nickel or cobalt would increase the moment, but hardly that the addition of nickel or cobalt would increase the moment in iron alloys; still less that the addition of rhodium, iridium or platinum would have a similar effect. The initial increases are shown in table 4.2 (where the tabulated values are prefixed by a minus sign, It may be of interest to give a table of alloys with the highest observed moments per atom (table 4.3).

Table 4.3. Iron alloys with high saturation magnetization.

	At.%	σ_0	q_f
Co	27·0	243	2·46
Ni	6·6	224	2·25
Rh	25·0	204	2·42
Ir	7·5	192	2·27
Pt	12·4	186	2·44

σ_0, specific saturation magnetization extrapolated to absolute zero ; q_f, corresponding mean number of parallel spins per atom. For iron, $\sigma_0=222$, $q_f=2{\cdot}22$.

The maximum in the mean moment, composition curve is usually fairly flat (for the iron-cobalt alloys, for example, the change is within 1 % from 22 to 34%). It may be noticed that although the q_f value is higher than that of iron for all the alloys, the specific magnetization is lower for the alloys with the non-ferromagnetic elements (the atomic weight being higher); it is only for the cobalt alloy that σ_0 is markedly higher (by about 10 %). For this alloy the volume intensity of magnetization is some 12 % greater than for iron, of which application is made in the use of ferrocobalt for the pole pieces of electromagnets; for the Fe–Ir and Fe–Pt alloys it is about the same as for iron. For the nickel-manganese alloy the maximum occurs at about 8·4 atomic

per cent Mn, the σ_0 and q_t values being 73·2 and 0·76 as compared with 57·6 and 0·61 for nickel. For further studies of this interesting alloy system reference may be made to the work of Valentiner and Becker (1935) and of Thompson (1940).

Electronic specific heat. In the expression (4.5) for the specific heat at very low temperatures, the γT term corresponds to an electronic contribution, and according to the usual theory, γ may be assumed to be proportional to the density of states at the top of the Fermi distribution. If the change in moment on the addition of copper to nickel is correctly interpreted as corresponding to a filling up of the holes in the d band, and if, further, the form of the d band at its upper end is anything like that represented in figure 3.2, it would be expected that in nickel-copper alloys, γ would decrease regularly from the high value for nickel ($\gamma \times 10^4 = 17{\cdot}44$) to a value not greatly different from the low value for copper ($\gamma \times 10^4 = 1{\cdot}78$) at about 60 % Cu. Measurements have been made by Keesom and Kurrelmeyer (1940) on a series of alloys of nickel with copper and also with iron. The values deduced for γ are shown in figure 4.4.

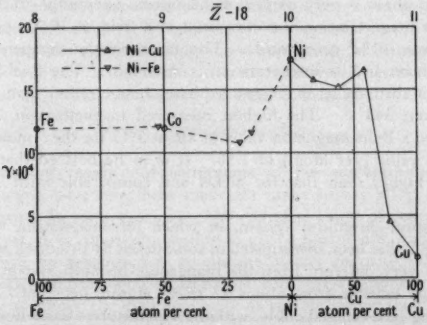


Figure 4.4. Values of γ for alloys of nickel with copper and with iron. γ, coefficient of T in the expression for the gram atomic heat at low temperatures. $\bar{Z}$, mean atomic number. (After *Keesom and Kurrelmeyer*, 1940.)

As Keesom and Kurrelmeyer point out, the curve for the nickel-copper alloys indicated by the experimental points is quite different from that predicted by the theoretical work of Mott and Slater; moreover, on the basis of these results, it is not easy to understand why the ferromagnetic Curie point and the saturation magnetization go linearly to zero at 60% copper. Even if the "end point" for ferromagnetism is less sharp than linear extrapolation of the earlier σ_0 and θ values would indicate (as is indeed suggested by the paramagnetic measurements referred to below) the band form implied by the specific heat results is very peculiar. Little further can be said at present about either the nickel-copper or the nickel-iron results. Work of this kind, however, is undoubtedly capable of providing most important information about energy band forms. This information must be linked up with that provided by experimental work on x-ray emission and absorption spectra, particularly in the long x-ray region (Farineau, 1938; Beeman and Friedmann, 1939; Skinner, 1939, 40; Friedmann

and Beeman, 1940), and by theoretical work on electronic energy bands in metals and alloys (see § 3 (iv)).

Alloys of manganese and chromium. Among alloys and compounds not containing a ferromagnetic element, ferromagnetism occurs for a considerable number containing manganese and a few containing chromium. Thus Mn forms ferromagnetic substances with all the fifth column elements N, P, As, Sb and Bi; and Cr with the sixth column elements S (Bates and Taylor, 1938) and Te. Partly because the work on these substances for the most part antedates the period mainly under consideration here, and has often been reviewed, and partly because the quantitatively indefinite character of many of the results would make a short summarizing statement of doubtful value, the bulk of this work must be passed over. In view of the interest of these ferromagnetic compounds of non-ferromagnetics, however, brief reference should perhaps be made to two fairly recent investigations. A detailed study of the manganese-bismuth system has been made by Thielmann (1940), who concludes that the ferromagnetic phase is very narrow and consists essentially of the compound MnBi. The magnetization was measured in a field of 10000 oersted at temperatures from $-192°$ c. upwards. The magnetization-temperature curve is of a curious form and its interpretation is uncertain. The final drop towards zero occurs so abruptly as to suggest a phase change rather than a Curie point proper at about 343° c. The highest measured magnetization, at $-180°$ c., corresponds to a Bohr magneton value of about 3·13 for the "molecule" MnBi, or a mean q_f value (per atom) of 1·56. It is to be noticed that this value is considerably higher than that for nickel and comparable with the value for cobalt.

The platinum-chromium system, in which ferromagnetism was found by Friedrich (1932), has been investigated in some detail by Friedrich and Kussmann (1935). It is very different from the manganese-bismuth system. The ferromagnetic range extends from about 22 to 48 atomic per cent Cr, the crystal structure being face-centred cubic, without superlattice except towards the end of the range. The magnetic behaviour is remarkable in that the Curie point rises steadily, from about 50° c. at 22% to 780° at 45%, while the saturation magnetization rises to a flat maximum at about 30% and then falls. The magnetization, in the vicinity of the maximum, is some three times as great for the alloy in the annealed state as when quenched from 1400° c. The highest observed value of $4\pi I_0$ was 3100, corresponding to a σ_0 value of about 14·5, or a mean q_f value per atom of about 0·4.

From a theoretical point of view very little more can be said now about the ferromagnetic binary alloys or compounds of Mn and Cr, or about the ternary alloys of Mn (of which the Heusler alloys, with optimum composition at $MnAlCu_2$, are the prototypes), than could be said some ten years ago. The necessary requirements for ferromagnetism are incomplete groups of electrons and positive interchange interaction; and both theory and experiment show that the last is dependent on the ratio of the interatomic distance to the effective radius of the incomplete electron group in the atom. The Mn atom has one less electron than iron, and with suitable neighbours in alloys, it may well become "iron-like". With a little ingenuity some of the particular facts could no doubt be

plausibly accounted for; but no really satisfactory theoretical treatment has yet been put forward. In only one respect has the situation been somewhat simplified, namely in that it need no longer be assumed that the absence of ferromagnetism necessarily entails the absence of positive interchange interaction (see §3(iv)). The transition from the paramagnetism of the metal manganese, which is so high that positive interchange interaction is almost certainly involved, to the ferromagnetism of many of the manganese alloys, does not imply nearly so drastic a modification in electronic structure as was previously supposed, and it may not be long before the magnetic properties of the critically placed element, manganese, alone and in combination, are covered by a more adequate co-ordinating scheme.

In this section, surveying generally some of the more recent work on the saturation magnetization of alloys, there is inevitably a lack of detail about particular alloy systems, and the details omitted may contain the key to some of the problems raised. Fuller details, and further references may, however, readily be found from the papers to which explicit reference is made. For the earlier work reference may first be made to the books listed in the bibliography. Among review articles not previously mentioned those of Auwers (1937, 38) are particularly relevant as containing a comprehensive analysis of the experimental material with which this section has been concerned.

4(iv). *Paramagnetic alloys*

General.

Representation of results. Measurements of the susceptibility over wide temperature ranges have now been made for many binary alloys of the ferromagnetic elements, in some cases over the complete range of composition. At present, however, it is almost impossible to give a useful summary of the data, and all that will be attempted is an outline of the character of the results, illustrated by material from some of the more recent investigations. It is believed that, as theory advances, the data may be of increasing importance as a means of gaining information about the electronic structure of alloys, and for this reason, even though most of the work is not reviewed, sufficient references are given to enable further details to be found if they are required. An indication of the reasons for the difficulty in summarizing the results will serve at the same time as an introduction to their general character.

If the susceptibility of an alloy followed a Curie-Weiss law,

$$\chi = C/(T-\theta_p), \qquad \ldots\ldots(4.6)$$

the magnetic behaviour could be summarized in the values of C and θ_p, and two curves, giving the variation of these with composition, would suffice to show the magnetic characteristics of the entire alloy system. The parameters C and θ_p, moreover, would be theoretically significant, and from them could be derived information about the number or magnetic moments of the carriers, and about the magnitude of the interchange interaction effects. This is the kind of information which is required, and which it seems should be obtainable from the magnetic investigations; but it cannot be obtained by forcing a Curie-Weiss law on recalcitrant data. For those alloys which are ferromagnetic, in the sense of having a definite ferromagnetic Curie temperature, θ_f, above absolute zero,

the susceptibility variation may be similar to that of the ferromagnetic elements (see figures 2.2, 4.2), and the representation and interpretation of the results may proceed along the lines already discussed (§4(ii)). Even for the ferromagnetic elements, however, where θ_f and θ_p are not very different, and where there is linearity over an extensive range in the $1/\chi$, T representation (as for Ni), there are considerable doubts as to whether the experimental Curie constant has the same relation, save as a rough approximation, to the number of effective spins per atom as is attributed to it by the simple theory. For the alloys these doubts become much greater, for in general the curvature of the $1/\chi$, T lines above θ_f is much more extended, an approximately linear range often not being reached, if at all, until the temperature is some 100° or 150° above θ_f or θ_p. The deviation from linearity at the higher temperature end is more often than not so marked that, even excluding the Curie point region, a two parameter representation of the results must be abandoned. The equation then usually tried is

$$\chi-\alpha=C/(T-\theta_p), \qquad \ldots\ldots(4.7)$$

where α, for representational purposes an adjustable constant, may be interpreted, if positive, as a superposed constant paramagnetism. (It has long been customary to apply a correction for diamagnetism in this way, with α negative). Theoretically, the α term is entirely justifiable, but to straighten the experimental curves (usually concave downward at the upper end) values of α are required which are in most cases considerably greater than can be accounted for by the known sources of constant paramagnetism (e.g. collective s electrons). Abandoning all attempts at a theoretically based equation, Kaufmann and Starr (1943), finding that neither (4.6) nor (4.7) fit the results for some of the copper-nickel alloys with less than 40% Ni, have used the frankly empirical equation,

$$\chi=aT+b+c/T. \qquad \ldots\ldots(4.8)$$

An equation of this type is required to cover cases in which the susceptibility increases with increasing temperature.

It may be mentioned that, with a suitable choice of α in (4.7), curves have been obtained for $1/(\chi-\alpha)$ for nickel-copper alloys (Gustafsson, 1937) and for nickel (Fallot, 1944) which, over a wide range including most of the Curie point region, are similar to those obtained theoretically by applying Fermi-Dirac statistics to electrons in a band of parabolic form (see figure 3.3). A proper examination of the bearing of the collective electron treatment on the results for alloys, however, has yet to be made. This may remove some of the present difficulties, but certainly not all of them. It is particularly to be noted that although this treatment gives a non-linear relation between $1/\chi$ and T (except in the classical limit), the relation would often be so nearly linear over moderate temperature ranges as to be of the Curie-Weiss type, but with θ_p having no direct significance, and with a Curie constant greater than that given by the quasi-classical treatment for the same number of electron spins per atom. When the experimental Curie constants are interpreted in the usual way, they often imply numbers of spins per atom which are unacceptably high, so the modifications introduced by the collective electron treatment will be, in this respect, in the right direction.

From the above discussion it will be clear that a mere list of experimental C and θ_p values, such as could be given, would be not only inadequate as a representation of the experimental results, but also liable to be misleading in the implications as to theoretical significance.

Classification of alloys. Following, with slight modifications, a scheme adopted by Auwers (1937), Néel (1940) and others, it is probably convenient to regard the alloys of ferromagnetic elements as falling, magnetically, into two broad classes. (In the present connection it is sufficient to consider only single-phase alloys.) First, there are the alloys with ferromagnetic or strongly paramagnetic metals, which have characteristics which may be attributed to an incomplete d group of electrons (e.g. Fe, Co, Ni; Pd, Pt; Mn); second, the alloys with diamagnetic, or feebly paramagnetic metals, in which the underlying groups or sub-groups are complete, and in which the paramagnetism, constant, or only slightly varying, and comparable in magnitude with the always present small diamagnetism, can be attributed to outer incomplete groups, usually of s or p electrons (e.g. Cu, Ag, Au; Zn; Al). The magnetic characteristics of alloys depend, of course, not only on the electronic structure of the constituent elements, but also on the magnitude of the interchange interaction between unlike as well as like pairs of atoms. Far too little is known quantitatively about interaction effects, however, to provide a basis for a more elaborate classification of the alloys, though it is not unreasonable to expect, in the second class above, widely different magnetic characteristics of dilute solid solutions according as the solute is the ferromagnetic or the non-ferromagnetic element. A few particular alloy systems will now be considered, one in some detail and the rest very briefly.

Nickel-copper. Following earlier work by Alder (1916), measurements over extended temperature ranges of the susceptibility of alloys of the nickel-copper series have been made in recent years by Gustafsson (1937, 0–88% Cu, 300–700° K.), Wheeler (1939, 0–60 Cu, 300–1000° K.) and Kaufmann and Starr (1943, 50–100 Cu, 14–1270° K.). Where there is an overlap in the temperature and composition range examined by different investigators, there may not be good quantitative argument in the values deducible for a particular alloy at a particular temperature; but there is agreement as to the general character of the results, which is conveyed by the curves of figures 4.5(*a*) and 4.5(*b*), drawn from the most recent data.

In these figures the full curves have been drawn smoothly through the experimental points, which show no systematic deviations. In the second figure, the experimental (full) curves are continued by a broken curve to a point on the axis corresponding to the (ferromagnetic) Curie temperature derivable from the results of Marian (1937). The points through which the curves have been drawn correspond to the inverse of the measured mass susceptibility without any correction. A correction for the underlying diamagnetism of the copper and nickel would not appreciably affect the curves shown in the figures, though such corrections would become important for lower nickel contents. It is obvious from the curves that a Curie-Weiss law (4.6) is in general inadequate. It has been seen (§4(ii)) that this law does hold over a wide range for pure nickel, but in the alloys the approximately linear range decreases as the copper content

increases. A better representation is given over the higher temperature ranges by the equation (4.7), the constant paramagnetism parameter α being treated as adjustable. From the curves of figure 4.5 (*b*), by suitable choice of α in each case, curves of the inverse of $\chi-\alpha$ can be obtained in which the downward concavity disappears, the curves being concave upwards in the lower part (towards the Curie point) and almost linear in the upper part. This procedure cannot, however, be applied with success to the curves of figure 4.5 (*a*). In particular, a susceptibility increasing with temperature (as shown by the 20·3 and 30·0 alloys) obviously cannot be accounted for by a superposed constant paramagnetism. The empirical formula (4.8) can be used, but the physical significance of the coefficients is very uncertain. It is not surprising that if c in this formula is treated as a Curie constant the Bohr magneton values p_B or the electron spin values q_p derived from it do not appear to fit into any self-consistent and significant scheme.

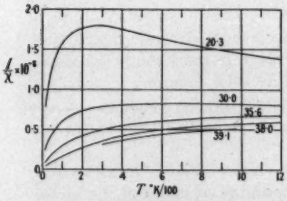


Figure 4.5 (*a*). Susceptibility of copper-nickel alloys with less than 40% nickel. The weight per cent nickel is shown on the curves. *Data from Kaufmann and Starr* (1943) except for 39·1 Cu, *from Wheeler* (1939).

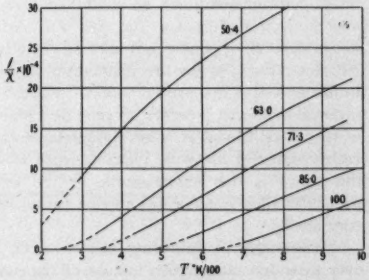


Figure 4.5 (*b*). Susceptibility of copper-nickel alloys with more than 40% nickel. *Data from Wheeler* (1939), except for pure nickel, *from Fallot* (1944).

Nickel-copper is one of the simplest alloy systems in that the alloys are single phase over the whole range of composition and temperature. Moreover, the Curie temperature of nickel is relatively low, so that the T/θ range which can be examined for the ferromagnetic alloys is fairly extensive. Thus nickel-copper is one of the most favourable alloy systems for magnetic investigation. Simple though the system is, the experimental results are far from having been accounted for in quantitative detail, and the above discussion shows how misleading an oversimplified representation of these results may be. In general, significant Curie constants cannot be unambiguously derived for the alloys, as is almost obvious when the results as a whole are considered, rather than the results for one alloy, or a pure metal. This being so, Curie constants given for other alloys, for which the effective temperature range examined may be much smaller, must be accepted only with the greatest reserve. It is for this reason that values which have been given for Curie constants are usually not quoted here. It is not, of course, implied that there are no intelligible regularities

in the results. If the series of curves in figure 4.5 (*b*) are examined it will be seen that the slopes (particularly if taken at temperatures correspondingly related to the Curie points) increase regularly as the nickel content decreases; that is, the effective Curie constant, given by $dT/d(1/\chi)$, decreases. The decrease runs parallel with the decrease in the effective number of spins per atom as derived from the low temperature saturation magnetization. The decrease, however, is not quantitatively in accordance with that indicated by the simple theory, though it is of the same order. A further point which may be noted is that there is no abrupt change in the character of the curves as the nickel content is changed from above to below 40%, though it is here that the low temperature results suggest a filling up of the holes in the d band of nickel and a change over from ferro- to paramagnetism. It may be suggested that the fundamental defect of the theoretical treatments which have hitherto been applied is that they are based on Maxwell-Boltzmann rather than on Fermi-Dirac statistics. The peculiarities in the results at the higher temperatures may be eventually explained on the lines indicated in the discussion of pure nickel (§4(ii)).

Other nickel alloys. The most extensive investigation of the susceptibilities of alloys of nickel is that of Manders (1936) who made measurements between room temperature and 710° c. on some eighty alloys in the solid solution ranges of the following eighteen elements in nickel: Al, Si; Ti, V, Cr, Mn; As; Zr, Mo, Ru, Pd; Ag, Sn, Sb; W, Pt; Au, Pb. The results could usually be represented with fair approximation at least over the upper part of the temperature range by (4.7). Suitable methods for deriving the "best" values of α and θ_p are given by Néel (1936). He finds that the relation between the inverse of $(\chi-\alpha)$ and T often does not become linear until some 100° or 150° above the Curie point, so that the temperature range of linearity given by these experiments is usually insufficiently extended for great weight to be placed on the derived values of the Curie constants. In a few cases it is possible to estimate a value for the initial change in the Curie constant per substituted atom. As examples the following values are found for $-\Delta C_A$ per substituted atom: Cu, 0·43; Al, 1·4; Sn, 2·3; Sb, 2·6. With a mechanism similar to that which covers fairly satisfactorily the effect of substituted atoms on the saturation magnetization (see table 4.2) it might be expected that the reduction of the Curie constant would be proportional to 1, 3, 4, 5, as very roughly it is; the mean proportionality factor, however, is rather greater than 0·5, as compared with the value 0·375 given by the simple theory (see equation (4.4)). Little purpose would be served by a fuller account here of the results of Manders, but his work is clearly a most valuable contribution, which provides extensive magnetic data on alloys which call for much further study. Apart from the alloys with Cu, already considered, other alloys of nickel investigated include those with Zn (Wheeler, 1939), and with Au (Kaufmann, Pan and Clark, 1945) and the inter-ferromagnetic alloys with Fe (Peschard, 1925) and Co (Bloch, 1912). Nickel amalgams with up to 2·68 % Ni have been studied by Bates and Baker (1940). Freshly prepared, these amalgams are diamagnetic, though increasingly more paramagnetic than mercury, with increasing Ni content, but on heating to about 224° c. they become ferromagnetic, in the sense that the field susceptibility curves show the presence of a constituent whose magnetization reaches a saturation value. The

ferromagnetism persists on cooling again. The appearance of ferromagnetism suggests a clustering of nickel atoms or ions. The study of amalgams of ferromagnetic metals might have been expected to give fairly direct information about some of the fundamental questions of ferromagnetism, but in detail the behaviour of these amalgams is surprisingly complicated, and raises further special problems, both of experimental technique and theoretical interpretation, for an account of which the original papers must be consulted.

Iron alloys. Among iron alloys whose paramagnetic properties have been studied, usually over more restricted composition and effective temperature ranges than the nickel alloys, are those with Al (Fallot, 1944), Si (Néel, 1932); V (Sucksmith and Pearce, 1938); Cu (Bitter and Kaufmann, 1939; Bitter, Kaufmann, Starr and Pan, 1941); Sn (Néel, 1932); Au (Pan, Kaufmann and Bitter, 1942; Kaufmann, Pan and Clark, 1945); Hg (Bates and Illsley, 1937). Mention has already been made of the effect of Al and V in "suppressing" the γ-phase of iron in the account of the paramagnetism of the element (§4(ii)). The values for the change in the Curie constant, where estimates have been made (for substitution of Al, Si, V and Sn), are not revealing; even the sign of the change is different for Si and Sn, both with four valence electrons (Néel, 1940). The equilibrium solid solubility of iron in copper at room temperature is very small, certainly less than 0·1%; according to the measurements of Bitter and Kaufmann, amounts of less than 0·001% may be shown up in the magnetization curves as a ferromagnetic constituent, which becomes saturated in fields of a few kilo-oersted. By quenching specimens from temperatures of about 1000° C., after annealing, up to 2 or 3 per cent iron can be obtained in solid solution, and such quenched alloys were studied in the 1941 work, measurements being made at temperatures from 140° to about 1300° K. The inverse of the susceptibility per gram of iron (obtained by dividing $\chi - \chi_{Cu}$ by the fractional concentration of Fe) may be plotted in the usual way against the temperature. At the lower temperatures (14° to 77° K.) good straight lines are obtained for iron percentages up to 0·7. The θ_p values are numerically small, and the Bohr magneton values per atom, p_B, increase with increasing iron content up to about 7; at the high temperatures, the slopes for the different alloys are about the same, the p_B values all being about 4·7, but the θ_p values vary with iron content. All that can be said is that these p_B values are of the order of magnitude found for iron ions in salts. No interpretation has been given of the precise values and there is no final explanation of such effects as the marked field dependence of the susceptibility found at the lower temperatures. In quenched alloys of this kind, the purely magnetic problems, which are still largely unsolved even for the most straightforward single phase alloys, are complicated by uncertainties as to whether a partial segregation of the iron atoms occurs. Magnetic studies of a different kind, of the hysteresis properties of alloys with low ferromagnetic contents, may assist towards the elucidation of the very complex problems of which the work under consideration illustrates one aspect.

Cobalt alloys. The paramagnetism of cobalt alloys has been comparatively little studied, the Co–Pt system being the only one, apart from the inter-ferromagnetic alloys, which has been investigated over a wide range of composition (Néel, 1932). For the ferromagnetic alloys (those with more than about 25

atomic per cent Co), the $1/\chi$, T curves become approximately linear some 150° above θ_f; derived Curie constants and θ_p values are given for these alloys and also for the non-ferromagnetic alloys with less Co, for some of which large negative θ values are obtained. Although some regularities emerge, the investigation must be regarded as essentially a preliminary magnetic survey of a very complex system. More recently solid solutions, obtained by quenching, of Co in Cu (up to 0·5 atomic per cent) and Co in Au (up to 3·4%) have been studied by Hildebrand (1937) in a temperature interval from −190° c. to +640° c. The part of the susceptibility which may be associated with the Co atoms follows a Curie-Weiss law, the moment per atom being approximately the same for different alloys with the same solvent metal, but different in Cu and Au (the p_B values are about 5·4 and 4·6 respectively). The θ_p values are large and negative (−860° to −1020° and −160° to −250° in the Cu–Co and Au–Co alloys respectively). These values are quoted to give some indication of orders of magnitude. The artificiality of this method of representation, except in special cases, has already been sufficiently stressed.

Other alloys. In a report on ferromagnetism it is appropriate to limit the consideration of paramagnetism to that of alloys which either are ferromagnetic over some part of the temperature range or contain a ferromagnetic element as a constituent. The change with composition from ferro- to paramagnetism at a particular temperature is not associated with an abrupt change in the paramagnetic properties at higher temperatures, and it would be entirely artificial to limit the composition range to be considered to that over which there is a definite positive ferromagnetic Curie temperature. Somewhat similarly no sharp differentiation can be made between paramagnetic alloys which contain a ferromagnetic element in solid solution, and those which contain other transition elements, particularly such elements as Mn and Cr and those of the palladium and platinum triads. The paramagnetism of a number of such alloy systems has been investigated, and although the results cannot be discussed here, reference may perhaps usefully be made to the fairly recent papers on alloys of Mn with Cu, Ag, Au (Gustafsson, 1936) and with Pd (Grube and Winkler, 1936), and of Pt with Cr (Friedrich and Kussmann, 1935) and with Rh (Hildebrand, 1937) and to a general paper by Vogt (1937), who has carried out much of the earlier work on metallic mixed crystals. This "footnote" on other alloys is added to this section because alloys containing non-ferromagnetic transition elements are probably better regarded as special cases of ferromagnetic alloys than conversely. The theoretical interpretation of the properties of ferromagnetic alloys and metals above the Curie point involves a treatment on the basis of Fermi-Dirac statistics, and a consideration of the effects of interchange interaction. In other alloys and metals the interchange interaction may still be positive, though relatively smaller, or it may be negative; only in very special cases is it likely to be strictly zero. Again, Maxwell-Boltzmann statistics, the basis of what has been called the "simple theory", is not the generally applicable form of statistics, to be replaced by Fermi-Dirac statistics only in special cases; it is rather a limiting form of Ferm-Dirac statistics, which may often be applied to give useful approximations, their usefulness, however, being largely dependent on an appreciation of the extent to which they are approximations.

4(v). *Transformations in alloys*

General. The form of the curve giving the variation with temperature of the quasi-saturation magnetization is, for many ferromagnetic alloys, much the same as that for the ferromagnetic metals (see figure 2.1), except for a more extended tail in the Curie point region, which increases the uncertainty in the determination of the Curie point. In the paramagnetic region, a smooth variation of χ, or its inverse, is also often found, as illustrated by the curves of figure 4.5 (*b*). Where there are more or less abrupt changes, or marked deviations from the "standard" type of variation with temperature, an explanation is usually to be sought in terms of phase changes or of other types of transformations in the solid state. The magnetic results on any particular alloy system must be considered in relation to the information as to constitution obtained by the standard metallurgical methods (particularly by metallographical examination and by thermal analysis) and as to the crystal structures present, and to some extent the character of the segregation, obtained by x-ray methods. It is important to bear in mind that the constitution of an alloy of given composition is not uniquely determined by the temperature; there may be marked constitutional differences between quenched and annealed alloys, for example, which are reflected in magnetic differences. The rate of transformation to a presumed equilibrium state may be a relevant factor resulting in a dependence of the form of a magnetization-temperature curve on the rate of heating or cooling. The underlying processes of which the magnetic changes are, as it were, a visible manifestation, are often complicated in detail, and very incompletely understood. In spite of their interest, a discussion of them here would be out of place. In what follows a few examples only will be given of some of the main types of magnetic behaviour associated with transformations, the examples being drawn from recent studies on two binary systems (Fe-Pt and Fe-Ni) and one ternary (Fe–Ni–Al).

Iron-platinum alloys, and the phase transformation diagram. A common type of transformation, both in metals and alloys, is that in which there is a change of crystal structure from, for example, body-centred cubic (which will be denoted by β) to face-centred cubic (which will be denoted by α). (This nomenclature is in accordance with recent practice. It conflicts with the historical use of γ for the face-centred phase of iron.). Provided that the rate of change of temperature is not too great, such transformations are usually completed over a small temperature range about a well defined mean value, but there is a temperature hysteresis, the change on heating (say $\beta \rightarrow \alpha$) occurring at a higher temperature than the converse change ($\alpha \rightarrow \beta$) on cooling. If the β phase is ferromagnetic, and the α phase non-ferromagnetic, or appreciably less ferromagnetic, a $\beta \rightarrow \alpha$ transformation occurring below the Curie temperature of the β phase will be shown magnetically by an abrupt decrease in the quasi-saturation magnetization (as measured in high fields). If such a change occurs above the Curie point for both phases, it will be indicated by an abrupt decrease in the susceptibility. A diagrammatic representation of the magnetic changes associated with such transformations is shown in figure 4.6. The temperatures denoted by T_1, T_1' correspond to the beginning and end of the transformation on heating, in a sense which will be clear from the diagram, and similarly for T_2, T_2' on cooling.

The type of transformation represented in figure 4.6 (*b*) occurs in iron, and also the converse change from a less to a more paramagnetic state (see figure 4.2, in which T_1 and T_2 for the two changes are not shown separately). Besides the two possibilities illustrated in figure 4.6, there is the third possibility $T_2 < \theta < T_1$, for which the normal Curie point change would occur on heating, but not on cooling. It is to be noted that whereas a lattice change is an irreversible phenomenon, the change from a ferromagnetic to a paramagnetic state, without a lattice change (that is, the normal Curie point change) is reversible. These

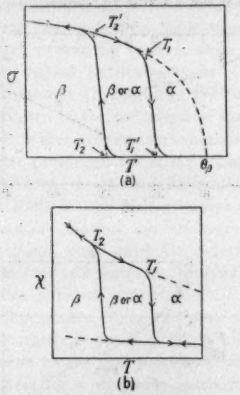


Figure 4.6. Diagrams illustrating the magnetic behaviour associated with phase changes.

β, ferromagnetic;
α, non-ferromagnetic.

(*a*) $T_2 < T_1 < \theta$.

(*b*) $\theta < T_2 < T_1$.

T_1, T_2, transformation temperatures on heating and cooling; θ, Curie temperature.

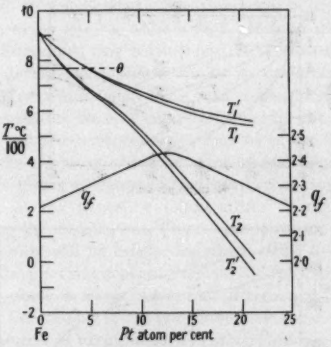


Figure 4.7. Transformation temperatures, Curie temperature and low temperature saturation moment of Fe-Pt alloys. The two lines for the transformation temperature on heating (T_1, T_1') and cooling (T_2, T_2') correspond to the beginning and end of the change. (See figure 4.6.)

β, body-centred cubic;
α, face-centred cubic.
q_f, saturation moment in Bohr magnetons per atom.
θ, Curie temperature.
(*After Fallot*, 1938.)

statements on alloy types, and on reversibility and irreversibility, are taken essentially from Fallot (1938), and, with the diagrams, are immediately relevant to the alloys of iron with the metals of the platinum and palladium triads, to his work on which reference has already been made. Besides measuring the low temperature saturation magnetization, Fallot also determined the Curie temperatures, θ, and the transformation temperatures of these alloys. For this purpose he made use of a convenient self-registering apparatus (Fallot, 1936); with the same apparatus, Curie points proper could be determined with good precision by using fields of 300 to 500 oersted, and transformation temperatures

with fields of some 5000 oersted, with a low sensitivity in the ferromagnetic and a high in the paramagnetic region. The transformation temperatures as determined by Fallot for Fe–Pt with up to 25 atomic per cent Pt are shown in figure 4.7. The low temperature (β phase) saturation moments (see table 4.2) are also shown. (The alloy at 25 Pt is face-centred, has a Curie temperature at 150° c., and shows no phase transformations. For more than 25 Pt, the alloys are two phase.)

Somewhat similar transformations occur, though over different ranges of composition, in alloys of iron with other elements of the Pd and Pt triads.

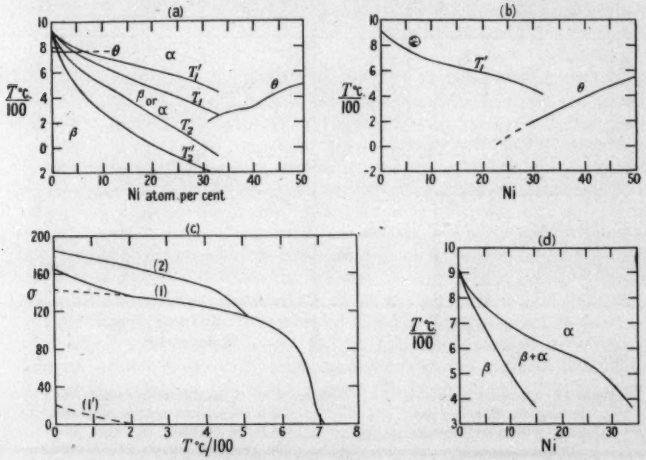


Figure 4.8. Transformations in iron-nickel alloys.

(*a*) Practical phase diagram, *after Peschard* (1925). θ, Curie temperature; T_1, T_2 transformation temperatures (see figure 4.6).

(*b*) Transformation temperatures, and Curie temperatures in pure single phase alloys.

(*c*) σ, T curves for 16% Ni alloys: (1) annealed 18 days at 525° c.; (2) annealed and subsequently cooled to −180° c. for 48 hrs.

(*d*) Equilibrium phase diagram of iron-rich Fe-Ni alloys.

(*b*) (*c*), (*d*), *after Sucksmith and Hoselitz* (1943). β, body-centred cubic; α, face-centred cubic.

Iron-nickel alloys and the phase equilibrium diagram. In the irreversible ferro-nickels (from 0 to 35% Ni) the behaviour, on which many investigations have been made, is again similar to that of the irreversible ferro-platinum alloys. The temperatures corresponding to the beginning and completion of the $\beta \rightarrow \alpha$ transformation on heating (T_1, T_1') and of the $\alpha \rightarrow \beta$ transformation on cooling (T_2, T_2') as determined by Peschard (1925) are shown in figure 4.8 (*a*). Although there are wide divergences in the various published curves (see, for example, Auwers, in *Gmelins Handbuch*, FeD, p. 129 (1938)), it is found by Sucksmith that with well quenched (pure single phase) alloys, and with

"normal" heating and cooling rates of 5° to 10° C. per min., the σ, T curves are repeatable to a high degree of coincidence, and that over most of the relevant composition range the temperature T_1', in particular, can be determined with considerable accuracy. A diagram such as figure 4.8 (*a*) for the alloys with up to 33% Ni, has been described as a "hysteretic" diagram. How it is related to the true equilibrium phase diagram is a question which can be answered only by fuller experimental investigation, such as has been carried out in an exhaustive magnetic study of these alloys by Sucksmith and Pickles (1940) and Sucksmith and Hoselitz (1943). Complete magnetization temperature curves were determined in fields usually of 16 kilo-oersted. Measurements of this kind, necessarily tedious, were made much less so, while a high precision was maintained, by development of the Sucksmith ring balance and the associated equipment (Sucksmith, 1939 a), a development which marks an important advance in the technique of magnetic measurements. The basic experimental fact is that over certain ranges of composition, the σ, T curves obtained for alloys after prolonged annealing within certain temperature ranges may be markedly different from those obtained with the alloys as quenched after several hours at 1000° C., the magnetization curves giving evidence of the presence of magnetically distinct phases segregated at the annealing temperature. Various methods were used for determining the equilibrium phase compositions at different temperatures, of which one only will be briefly described. The $\beta \rightarrow \alpha$ transformation temperatures, T_1', of the pure single phase β alloys, and the Curie temperatures, θ, of the pure single phase α alloys were first determined as accurately as possible (figure 4.8 (*b*)). For an alloy in the two phase state, the σ. T curve shows two transition temperatures, as exemplified in curve (1) of figure 4.8 (*c*) for a 16Ni alloy annealed (after the initial quench treatment) for 18 days at 525° C. One transition is shown by an inflexion in the curve, the second by the final fall to zero of the magnetization. The first corresponds to the Curie temperature of the Ni-rich α phase, the second to the $\beta \rightarrow \alpha$ transformation of the Fe-rich β phase, and from the temperatures the compositions of the two phases can be estimated from figure 4.8 (*b*). The composition of the Ni-rich phase can be more exactly determined by cooling the alloy to $-180°$ C. for a short period before examination. The phase undergoes an irreversible transformation to the β form, and in the σ, T curve then obtained (curve (2) of figure 4.8 (*c*)) the composition is obtained from the temperature of transition to the α form. The two methods of determining the composition are in agreement, but the second is more precise. The upper part of the σ, T curve is unchanged by cooling to the low temperature. By curves such as that of figure 4.8 (*c*), the phase boundaries at 525° C. are determined at 9½Ni and 26½Ni. These points are plotted in figure 4.8 (*d*). In practice, a series of alloys of different compositions are examined. This very elegant method is suitable for annealing temperatures from about 525° C. to as low as 365° C., at which the alloys were annealed for 150 days; it greatly extends in the downward direction the temperature range that can be examined. At the higher temperatures, whether or not an alloy is of composition corresponding to an equilibrium two-phase or single-phase region is determined by whether or not the σ, T curve of the quenched alloy is modified by annealing at the temperature. The resulting equilibrium diagram is shown in figure 4.8 (*d*).

The temperature T_1' corresponding to the completion of the $\beta \rightarrow \alpha$ change in the quenched single-phase alloys is the same as the temperature of equilibrium between the $\beta+\alpha$ and α regions, but neither the temperature T_2 nor T_2' associated with the $\alpha \rightarrow \beta$ change on cooling have any general clear-cut relation to the temperature corresponding to the β, $\beta+\alpha$ equilibrium boundary. The value of this magnetic investigation lies not so much in that it has led to a precise determination of a portion of an equilibrium diagram to which the previous x-ray work by Owen and Sully (1939) and Bradley and Goldschmidt (1939) had given fair approximations, as in that it has made clear the precise relation between the strict equilibrium states of the alloy system and those non-equilibrium, but very stable, states which are normally found in practice. Moreover, in parts of the work not discussed here, it is demonstrated how magnetic methods may be effectively used in the study of the mechanism and kinetics of phase segregation.

Iron-nickel-aluminium alloys.

Order-disorder and other transformations. Probably the most extensive investigation that has been carried out on the more fundamental magnetic properties of ternary alloys is that of Sucksmith (1939 b) on the iron-nickel-aluminium system. The magnetic investigations were made on the same specimens as were used in the previous exhaustive x-ray analysis of the system by Bradley and Taylor (1937, 38). Over eighty alloys were examined, σ, T curves being obtained in fields of from 9000 to 18000 oersted for both quenched and

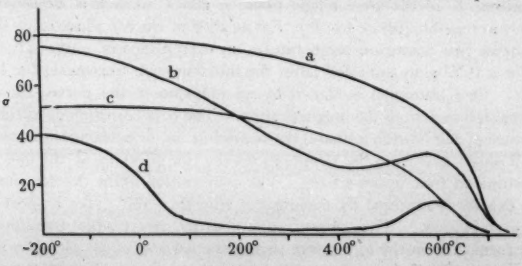


Figure 4.9. Magnetization, temperature curves for the ternary alloys 50Fe 20Ni 30Al (a, b), and 30Fe 30Ni 40Al (c, d). a, c, annealed; b, d, quenched. Annealed alloys, β_2; quenched alloys β changing to β_1. (Sucksmith, 1939.) (*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

annealed specimens (the usual procedure, after a preliminary lump anneal at 1300° C. for three days, was to quench in cold water from 900° C., and to anneal by cooling in vacuo from 900° C. at about 10° C. per hour). The estimated saturation magnetization, σ_0, at absolute zero, and the Curie temperature, or other transformation temperatures are given for all the alloys, as well as a considerable number of complete σ, T curves, illustrative of the magnetic properties in the various phase fields. For details of the many interesting results, and a discussion of them in relation to the room temperature ternary diagram as determined by x-ray examination of the annealed alloys reference must be made to the original

paper. In the light of the preceding discussion, the interpretation of some of the σ, T curves is fairly straightforward, though the curves may be more complicated in detail than those which have been explicitly considered. It must suffice to give here two sets of curves which illustrate how other types of transformation are shown up magnetically. In figure 4.9, curves are shown for two alloys which in the annealed state are single-phase body-centred cubic with a superlattice like FeAl (such a phase is specified by β_2). The quenched alloys, in which the atomic arrangement will be random, have, in these cases, a lower saturation intensity, and the initial parts of the curves b and d indicate that the Curie temperature associated with the random arrangement is lower than with the ordered arrangement; in the region 400–500° c., however, order begins to set in, the magnetization increases, and the curve finally joins that for the annealed alloy.

It may be noted that, owing to the difference in the magnetic characteristics of the ordered and disordered structures, an order-disorder transformation may be shown up magnetically in much the same way as a phase change or a ferromagnetic-paramagnetic transition, though effectively it is usually much less abrupt than the former and more abrupt than the latter. Remarkable sequences of changes are observed when, as sometimes happens, there is a superposition of the effects of diverse types of transformation.

As a final example of magnetic aspects of transformation in alloys, σ, T curves are shown in figure 4.10 for the 50Fe 25Ni 25Al alloy, which is of particular

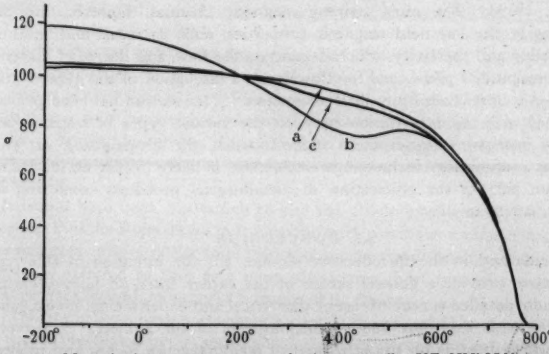


Figure 4.10. Magnetization, temperature curves for the ternary alloy 50Fe 25Ni 25Al. a, annealed; b, quenched from 800° c.; c, quenched from 500 °c. Annealed alloy, $\beta+\beta_2$. (Sucksmith, 1939.) (*Reproduced, by kind permission, from the 'Proceedings of the Royal Society'.*)

interest as the prototype of an important series of alloys which, after suitable heat treatment, have a very high coercivity, and are used for permanent magnets. Bradley first suggested on the basis of the x-ray work that the alloy is single-phase at high temperature, but splits into two body-centred phases (β and β_2) of composition approximately 95 Fe 2·5 Ni 2·5 Al and 30 Fe 35 Ni 35 Al, on slow cooling. Although the σ, T curve (a) for the annealed alloy appears to be of

the single-phase type, it is in fairly satisfactory agreement with the composite curve for the two phases presumed to be present. The quenched alloy, curve (c), has a higher saturation intensity than the annealed alloy; the form of the curve is interpreted as corresponding to a breaking up of the one phase state into two phases, beginning at about 400° C. (It would seem possible that ordering of the Ni-rich phase is also involved). The alloy quenched from 500° C., which shows intermediate characteristics, is in a state approximating to that associated with maximum coercive force.

Dispersion and precipitation phenomena. An alloy which is single-phase at high temperatures often remains single-phase on quenching to lower temperatures, although the equilibrium state at these temperatures is two-phase. Heat treatment at intermediate temperatures, or in some cases even ageing at the lower temperature, may result in a gradual transition to the two-phase state. The process of segregation, or precipitation, is often associated, particularly in the initial stages, by marked changes in the physical characteristics, such as the hardness, of the alloy, and the process is sometimes known as precipitation or dispersion hardening. For alloys in which at least one of the phases is ferromagnetic, the process is accompanied by changes in the magnetic properties. This is illustrated by figure 4.10, though the underlying process there is particularly complex. The somewhat more straightforward process occurring in Ni–Be, Ni–Au and other alloys has been examined by the magnetization temperature curve method by Gerlach (1936, 37 and 38). (See also Köster and Dannöhl, 1936; Becker and Döring, 1939). The most striking magnetic changes, however, are those occurring in the low field magnetic properties, such as initial and maximum permeability and coercivity, which depend on the form and degree of dispersion of the precipitated phase, and for this reason a discussion of this type of transformation is best taken up in that connection. This section has been primarily concerned with the interrelation between the various types of transformation and the instrinsic magnetization characteristics, the investigation of which, now that a satisfactory technique is established, is likely to play an increasingly important part in the elucidation of metallurgical problems connected with transformations in alloys.

§ 5. CONCLUSION

As indicated in the introductory section, § 1, the intention in this report has been to provide a general sketch of the earlier work on ferromagnetism, and a more detailed survey of recent theoretical and experimental investigations connected with the more fundamental aspects of the subject. A survey of investigations bearing on the behaviour of ferromagnetics in low and moderate fields has been deferred to a subsequent report. In this concluding section it may be appropriate to indicate the relation between this report and other review articles and books dealing with the same themes.

This report, like any report which covers a considerable period and which is intended to be reasonably self-contained, inevitably overlaps to some extent books and general articles written during the period under review. Thus, shorter or longer accounts of much of at least the earlier work discussed here may usually be found in one or other of the books listed in section A of the References. The list includes two books, those of Becker and Döring (1939) and of Bitter

(1937), concerned entirely with ferromagnetism, though in both the predominant themes are those connected with the magnetization curve aspect of the subject. More general books on magnetism, with extensive sections on ferromagnetism, are those of Bates (1939) and Stoner (1934, 1947). From the point of view presented in this report, the modern theory of the solid state, particularly the metallic state, provides the best approach to the theoretical understanding of ferromagnetism, and for this reason a number of books which deal with the general theory are included, namely those of Fowler and Guggenheim (1939), Mott and Jones (1936), Seitz (1940) and Wilson (1939). In all of these consideration is given to the special problem of ferromagnetism.

Among the topics not considered in this report, apart from incidental reference, is that of the ferromagnetism of substances which are not, in the usual sense, metallic, such as the oxides and sulphides of iron, and a considerable number of compounds with non-metals or metalloids of both ferromagnetic and non-ferromagnetic metals. These materials should provide an interesting, if difficult field for much further physico-chemical investigation. Knowledge of the magnetic properties, is, for the most part, both incomplete and lacking in precision; and what is known is little understood. Accounts of work on such substances are included in the books of Klemm (1936) and of Selwood (1943). Incidentally, Selwood's book is invaluable as a guide to the literature. Well over a thousand references are given, mainly to magnetochemical papers published over the period from 1934 to the end of 1942. A short discussion of ferromagnetism, in a metallurgical context, is given by Dehlinger (1939).

The character of *Gmelins Handbuch* is well known. As a systematic compilation of data and references the sections by Auwers (1938) dealing with the magnetic properties of iron and its alloys are invaluable. The reports and discussions at the Strasbourg Conference in 1939, published in *Le Magnétisme* (1940) give a good general impression of the outlook on magnetism just before the war. Among the reports, that of Néel, to which frequent reference has been made, is of particular value in connection with many of the matters discussed in §§ 3 and 4.

It would have been misleading to cite the article by Bozorth (1940), listed in section B of the References, in connection with particular topics in this report, as its aim is to cover a rather wide field without too much detail. It may appropriately be referred to here as a most stimulating short general review of the fundamental physical concepts of ferromagnetic theory.

Explicit citations have been made earlier in this report of some of the books and papers mentioned in this section, and of all the papers not mentioned here and which are included in the References. No attempt has been made at completeness in the references. In particular, papers in the nature of preliminary communications have been deliberately excluded. It is hoped, however, that the references for the period since 1934 are sufficient to enable original work in the field under review both during that period and earlier to be traced without undue difficulty.

ACKNOWLEDGMENT

It is a pleasure to acknowledge the assistance I have had from Mr. E. P. Wohlfarth in the preparation of this Report.

REFERENCES

References are not included for work published before 1934. For the earlier work mentioned in the Report, references may be found in Stoner, *Magnetism and Matter* (1934), or in other books listed in Section A.

A. Books

v. Auwers, O., 1938. *Magnetische und elektrische Eigenschaften des Eisens und seiner Legierungen. Aus Gmelins Handbuch der anorganischen Chemie.* (Berlin : Verlag Chemie.)

Bates, L. F., 1939. *Modern Magnetism.* (Cambridge: University Press.)

Becker, R. and Döring, W., 1939. *Ferromagnetismus.* (Berlin : Springer.) (Photo-Lithoprint Reproduction, 1943. Ann Arbor, Michigan : Edwards Brothers.)

Bitter, F., 1937. *Introduction to Ferromagnetism.* (New York : McGraw-Hill.)

Dehlinger, U., 1939. *Chemische Physik der Metalle und Legierungen.* (Leipzig : Akad. Verlags.)

Fowler, R. H. and Guggenheim, E. A., 1939. *Statistical Thermodynamics.* (Cambridge : University Press.)

Klemm, W., 1936. *Magnetochemie.* (Leipzig : Akad. Verlags.)

Le Magnétisme, 1940. Vol. I. *Généralités et Magnéto-Optique.* Vol. II. *Ferromagnétisme.* Vol. III. *Paramagnétisme.* (Report of an international conference held at Strasbourg, 1939.) (Paris : Institut International de Coopération Intellectuelle.)

Mott, N. F. and Jones, H., 1936. *The Theory of the Properties of Metals and Alloys.* (Oxford : Clarendon Press.)

Seitz, F., 1940. *The Modern Theory of Solids.* (New York : McGraw-Hill.)

Selwood, P. W., 1943. *Magnetochemistry.* (New York : Interscience Publishers.)

Stoner, E. C., 1934. *Magnetism and Matter.* (London : Methuen.)

Stoner, E. C., 1947. *Magnetism.* (Methuen's Monographs on Physical Subjects.) (London : Methuen.)

Wilson, A. H., 1939. *Semi-conductors and Metals.* (Cambridge Physical Tracts.) (Cambridge : University Press.)

B. Papers

The titles are given of papers which are primarily review articles.

v. Auwers, O., 1937 a. "Ferromagnetismus." *Ergeb. d. exakt. Naturwiss.*, **16**, 133.

v. Auwers, O., 1937 b. "Über einige Regeln der magnetischen Bindung." *Wiss. Veröff. a.d. Siemens-Werken*, **16**, 92.

v. Auwers, O., 1938. "Weitere Untersuchungen über die magnetische Bindung". *Wiss. Veröff. a.d. Siemens-Werken*, **17**, 74.

Awbery, J. H. and Griffiths, E., 1940. *Proc. Roy. Soc.* A, **174**, 1.

Band, W., 1946. *Proc. Camb. Phil. Soc.*, **42**, 139; 144.

Barnett, S. J., 1940. *Proc. Amer. Acad.*, **73**, 401.

Barnett, S. J., 1944. *Proc. Amer. Acad.*, **75**, 109.

Bates, L. F., 1937. "Magnetism." *Phys. Soc. Rep. Progr. Phys.*, **3**, 175.

Bates, L. F. and Baker, C. J. W., 1940. *Proc. Phys. Soc.*, **52**, 436; 443.

Bates, L. F. and Illsley, P. F., 1937. *Proc. Phys. Soc.*, **49**, 611.

Bates, L. F. and Taylor, G. C., 1938. *Proc. Phys. Soc.*, **51**, 33.

Beeman, W. W. and Friedmann, H., 1939. *Phys. Rev.*, **56**, 392.

Bitter, F. and Kaufmann, A. R., 1939. *Phys. Rev.*, **56**, 1044.

Bitter, F., Kaufmann, A. R., Starr, C. and Pan, S. T., 1941. *Phys. Rev.*, **60**, 134.

Bozorth, R. M., 1940. "The physical basis of ferromagnetism." *Bell System Tech. J.*, **19**, 1.

Bradley, A. J. and Goldschmidt, H. J., 1939. *J. Iron and Steel Inst.*, no. 2, p. 11 P.

Bradley, A. J. and Taylor, A., 1937. *Proc. Roy. Soc.* A, **159**, 56.

Bradley, A. J. and Taylor, A., 1938. *Proc. Roy. Soc.* A, **166**, 353.

Coeterier, F., 1935. *Helv. Phys. Acta*, **8**, 522.

Dehlinger, U., 1936. *Z. Metallkunde*, **28**, 116; 194.

Dehlinger, U., 1937. *Z. Metallkunde*, **29**, 388.

DUYCKAERTS, G., 1939. *Physica*, **6**, 401.
EDWARDS, OLIVE S. and LIPSON, H., 1943. *J. Inst. Metals*, **69**, 177.
FALLOT, M., 1936. *Ann. Phys., Paris*, **6**, 305.
FALLOT, M., 1937. *Ann. Phys., Paris*, **7**, 420.
FALLOT, M., 1938. *Ann. Phys., Paris*, **10**, 291.
FALLOT, M., 1944. *J. Phys. Radium*, **5**, 153.
FARCAS, T., 1937. *Ann. Phys., Paris*, **8**, 146.
FARINEAU, J., 1938. *Ann. Phys., Paris*, **10**, 20.
FAY, C. H., 1935. *Proc. Nat. Acad. Sci.*, **21**, 537.
FORRER, R., 1939. *J. Phys. Radium*, **10**, 181.
FORRER, R., 1940. "Ferromagnétisme: moment, spin et orbites; point de Curie et interactions orbitales." *Le Magnétisme*, Vol. II, p. 165. (See Section A.)
FRIEDERICH, E. and KUSSMANN, A., 1935. *Phys. Z.*, **36**, 185.
FRIEDMANN, H. and BEEMAN, W. W., 1940. *Phys. Rev.*, **58**, 400.
GERLACH, W., 1936. *Z. Metallkunde*, **28**, 80; 183.
GERLACH, W., 1937. *Z. Metallkunde*, **29**, 102; 124.
GERLACH, W., 1938. "Die Analyse der Ausscheidungshärtung mit ferromagnetischen Messungen." *Probleme der technischen Magnetisierungskurve* (ed. R. BECKER), p. 141. (Berlin: Springer.)
GERLACH, W., 1939. "Ferromagnetische Umwandlungen. Das Problem der Curie-Temperatur." *Z. f. Elektrochemie*, **45**, 151.
GREENE, J. B. and MANNING, M. F., 1943. *Phys. Rev.*, **63**, 203.
GRUBE, G. and WINKLER, O., 1936. *Z. f. Elektrochemie*, **42**, 815.
GUSTAFSSON, G., 1936. *Ann. Phys., Lpz.*, **25**, 545.
GUSTAFSSON, G., 1937. *Ann. Phys., Lpz.*, **28**, 121.
HILDEBRAND, E., 1937. *Ann. Phys., Lpz.*, **30**, 593.
HIRONE, T., 1938. *Sci. Rep. Tôhoku Univ.*, **27**, 101.
HOLSTEIN, T. and PRIMAKOFF, H., 1940. *Phys. Rev.*, **58**, 1098.
KAUFMANN, A. R. and STARR, C., 1943. *Phys. Rev.*, **63**, 445.
KAUFMANN, A. R., PAN, S. T., and CLARK, J. R., 1945. *Rev. Mod. Phys.*, **17**, 87.
KEESOM, W. H. and CLARK, C. W., 1935. *Physica*, **2**, 513.
KEESOM, W. H. and KURRELMEYER, B., 1939. *Physica*, **6**, 633.
KEESOM, W. H. and KURRELMEYER, B., 1940. *Physica*, **7**, 1003.
KLEMM, W. and BOMMER, H., 1937. *Z. anorg. u. allgemeine Chemie*, **231**, 138.
KÖSTER, W. and DANNÖHL, W., 1936. *Z. Metallkunde*, **28**, 248.
KRAMERS, H. A., 1936. *Leiden Comm.* Sup. 83.
KRUTTER, H. M., 1935. *Phys. Rev.*, **48**, 664.
KUSSMANN, A. and SCHULZE, A., 1937. *Phys. Z.*, **38**, 42.
LUDLOFF, H., 1934. *Z. Phys.*, **91**, 742.
MCDOUGALL, J. and STONER, E. C., 1938. *Phil. Trans. Roy. Soc.* A, **237**, 67.
MANDERS, C., 1936. *Ann. Phys., Paris*, **5**, 167.
MANNING, M. F., 1943. *Phys. Rev.*, **63**, 190.
MARIAN, V., 1937. *Ann. Phys., Paris*, **7**, 459.
MOTT, N. F., 1935. *Proc. Phys. Soc.*, **47**, 571.
MOTT, N. F. and JONES, H., 1937. *Proc. Roy. Soc.* A, **162**, 49.
NÉEL, L., 1934. *J. Phys. Radium*, **5**, 104.
NÉEL, L., 1935. *J. Phys. Radium*, **6**, 27.
NÉEL, L., 1936. *Ann. Phys., Paris*, **5**, 232.
NÉEL, L., 1937. *Ann. Phys., Paris*, **8**, 237.
NÉEL, L., 1940. "Champ moléculaire, aimantation à saturation et constantes de Curie des éléments de transition et de leurs alliages." *Le Magnétisme*, Vol. II, p. 65. (See Section A.)
NIESSEN, K. F., 1939. *Physica*, **6**, 1011.
OERTEL, L., 1937. *Phys. Z.*, **107**, 746.
OPECHOWSKI, W., 1937. *Physica*, **4**, 181; 715.
OWEN, E. A. and SULLY, A. H., 1939. *Phil. Mag.*, **27**, 614.
PAN, S. T., KAUFMANN, A. R. and BITTER, F., 1942. *J. Chem. Phys.*, **10**, 318.
POTTER, H. H., 1934. *Proc. Roy. Soc.* A, **146**, 362.
POTTER, H. H., 1937. *Proc. Phys. Soc.*, **49**, 671.

RADO, G. T. and KAUFMANN, A. R., 1941. *Phys. Rev.*, **60**, 336.
SKINNER, H. W. B., 1939. "The soft X-ray spectroscopy of the solid state." *Phys. Soc. Rep. Progr. Phys.*, **5**, 257.
SKINNER, H. W. B., 1940. *Phil. Trans. Roy. Soc.* A, **239**, 95.
SLATER, J. C., 1936. *Phys. Rev.*, **49**, 537; 931.
SLATER, J. C., 1937. *Phys. Rev.*, **52**, 198.
STEINHAUS, W., KUSSMANN, A. and SCHOEN, E., 1937. *Phys. Z.*, **38**, 777.
STONER, E. C., 1935. *Phil. Mag.*, **19**, 565.
STONER, E. C., 1936 a. *Proc. Roy. Soc.* A, **154**, 656.
STONER, E. C., 1936 b. *Phil. Trans. Roy. Soc.* A, **235**, 165.
STONER, E. C., 1936 c. *Phil. Mag.*, **22**, 81.
STONER, E. C., 1937. *Phil. Mag.*, **23**, 833.
STONER, E. C., 1938 a. *Proc. Roy. Soc.* A, **165**, 372.
STONER, E. C., 1938 b. *Phil. Mag.*, **25**, 899.
STONER, E. C., 1938 c. *Proc. Leeds Phil. Soc.*, **3**, 457.
STONER, E. C., 1939 a. *Proc. Roy. Soc.* A, **169**, 339.
STONER, E. C., 1939 b. *Phil. Mag.*, **28**, 257.
SUCKSMITH, W., 1935. *Helv. Phys. Acta*, **8**, 205.
SUCKSMITH, W., 1939 a. *Proc. Roy. Soc.* A, **170**, 551.
SUCKSMITH, W., 1939 b. *Proc. Roy. Soc.* A, **171**, 525.
SUCKSMITH, W. and HOSELITZ, K., 1943. *Proc. Roy. Soc.* A, **181**, 303.
SUCKSMITH, W. and PEARCE, R. R., 1938. *Proc. Roy. Soc.* A, **167**, 189.
SUCKSMITH, W. and PICKLES, A. T., 1940. *Proc. Roy. Soc.* A, **175**, 331.
SYKES, C. and WILKINSON, H., 1938. *Proc. Phys. Soc.*, **50**, 834.
THIELMANN, K., 1940. *Ann. Phys., Lpz.*, **37**, 41.
THOMPSON, N., 1940. *Proc. Phys. Soc.*, **52**, 217.
TROMBE, F., 1937. *Ann. Phys., Paris*, **7**, 385.
URBAIN, G., WEISS, P. and TROMBE, F., 1935. *C.R. Acad. Sci., Paris*, **200**, 2132.
VALENTINER, S., 1935. *Z. Phys.*, **97**, 745.
VALENTINER, S. and BECKER, G., 1935. *Z. Phys.*, **93**, 795.
VAN VLECK, J. H., 1936. *Phys. Rev.*, **49**, 232.
VAN VLECK, J. H., 1945. "A survey of the theory of ferromagnetism." *Rev. Mod. Phys.*, **17**, 27.
VOGT, E., 1937. *Ann. Phys., Lpz.*, **29**, 358.
WHEELER, M. A., 1939. *Phys. Rev.*, **56**, 1137.

THE CALCULATION OF ATOMIC STRUCTURES

By D. R. HARTREE, F.R.S.,
Cavendish Laboratory, Cambridge

§1. INTRODUCTION

THE structure of an atom as a whole has two aspects, the structure of the nucleus of protons and neutrons, and the structure of the extra-nuclear system of electrons; and to a good approximation these structures can be considered independently. This article is concerned entirely with the calculation of the extra-nuclear electronic structure; the nucleus of an atom, of atomic number N, will be considered simply as a point charge $N\boldsymbol{e}$, and, unless the contrary is specified, of infinite mass.

In considering the electronic structure, it is convenient to use the following natural atomic units:—

Unit of charge $= \boldsymbol{e}$ = magnitude of charge on the electron;
Unit of mass $= \boldsymbol{m}_e$ = rest mass of electron;
Unit of length $= a_H = \boldsymbol{h}^2/4\pi\boldsymbol{m}_e\boldsymbol{e}^2$ = radius of first Bohr orbit of the hydrogen atom.

Consistently with these, we have

Unit of energy $= \boldsymbol{e}^2/a_H = 2 \times$ (ionization energy of the hydrogen atom);
Unit of velocity $= 2\pi\boldsymbol{e}^2/\boldsymbol{h} = \boldsymbol{c}/137$;

and the value of $\boldsymbol{h}/2\pi$ is unity; the corresponding units of a number of other physical quantities are given in Condon and Shortley (1935), p. 432. These units will be taken as understood.

The problem of calculating the atomic structure of an atom or ion, other than one with just a single electron, is the problem of solving the wave equation for a many-particle system. This equation has no exact formal solution in finite terms, and it is necessary to use approximations both for this reason and for a more practical one, which is this: for an atom of p electrons, the solution is a function of $3p$ variables, and even if it were possible to evaluate such a solution to any degree of numerical accuracy required, no satisfactory way of presenting the results, either in tabular or graphical form, is known. It has been said that the tabulation of a function of one variable requires a page, of two variables a volume, and of three variables a library; but the full specification of a single wave function of neutral Fe is a function of seventy-eight variables. It would be rather crude to restrict to ten the number of values of each variable at which to tabulate this function, but even so, full tabulation of it would require 10^{78} entries, and even if this number could be reduced somewhat from considerations of symmetry, there would still not be enough atoms in the whole solar system to provide the material for printing such a table.

Hence in order to have wave functions in forms which it is practicable to calculate, to tabulate, and to manipulate so as to derive other atomic properties, it is quite necessary to have recourse to approximation. The approximation which has

been most widely used is one based on the idea of regarding each single electron as in a stationary state in the field of the nucleus and the space-charge distribution of the other electrons. Such a concept is of course much older than the present form of quantum mechanics; it is implicit in the first attempts to apply the concepts of Bohr's theory to atoms other than hydrogen, and is seen particularly clearly in the early interpretation of X-ray spectra as due to transitions of a single inner electron between various stationary states, and in Bohr's treatment of the periodic system of the elements.

This concept is very conveniently expressed in wave mechanics by approximating to the wave function Ψ of the whole atom by a combination of functions of the coordinates of the individual electrons. It is convenient to write these one-electron wave functions in the form $\psi(\alpha; j)$, $\psi(\beta; k)\ldots$, where $\alpha, \beta\ldots$ are labels distinguishing the different functions, and $j, k\ldots$ stand for the coordinates (space and spin) of the electrons $j, k\ldots$; the set of wave functions $\alpha, \beta\ldots$ will be written A, and the set of electron coordinates will be written R. Further, Pauli's principle is very conveniently expressed by combining these one-electron wave functions into a determinant:—

$$\Psi(\mathrm{A};\mathrm{R}) = \begin{vmatrix} \psi(\alpha;1) & \psi(\beta;1)\ldots\psi(\pi;1) \\ \psi(\alpha;2) & \psi(\beta;2)\ldots\psi(\pi;2) \\ \cdots\cdots\cdots\cdots \\ \psi(\alpha;p) & \psi(\beta;p)\ldots\psi(\pi;p) \end{vmatrix}. \qquad \ldots\ldots(1)$$

In applying this approximation, it is usual to neglect spin-orbit interaction in determining the one-electron wave functions ψ, and to take these as being of the form of the wave functions of an electron in a central field, viz.:—

$$\psi(\alpha;j) = \frac{1}{r_j}\mathrm{P}(n_\alpha, l_\alpha; r_j)\mathrm{S}(l_\alpha,(m_l)_\alpha;\ \theta_j, \phi_j)\chi((m_s)_\alpha;\ s_j), \qquad \ldots\ldots(2)$$

where $(n_\alpha, l_\alpha, (m_l)_\alpha, (m_s)_\alpha)$ are the quantum numbers of the wave function α and $(r_j, \theta_j, \phi_j, s_j)$ the spherical polar and spin coordinates of electron j; S is a spherical harmonic and χ a spin wave function. The relation between this choice of one-electron wave functions and the spatial symmetry properties of the atom as a whole has been discussed by Delbrück (1930, §2(2)) and Stevenson (1937). Wave functions of this form which differ in any one (or more) of the quantum numbers (l, m_l, m_s) are orthogonal through the properties of the spherical harmonic and spin factors. It is no restriction to assume that those for which (l, m_l, m_s) are the same are also mutually orthogonal, since if they were not, orthogonal linear combinations could be made without altering the wave function Ψ of the whole system. The further analysis is much simplified if all the one-electron wave functions ψ in the determinant (1) are orthogonal, and it will be assumed that they have this property.

In a central field there are $2(2l+1)$ wave functions (2) with specified values of (nl). A set of wave functions with the same (nl) is referred to as a *group* (sometimes as a *shell*, but this term is also used for a set of wave functions with the same n, and is best restricted to this sense); if all $2(2l+1)$ of these wave functions occur in the determinant (1), the group is called *closed* or *complete*. The set of numbers, $q(nl)$, of occupied wave functions in the different groups is said to specify a *con-*

figuration of the electronic system; the values of l are usually indicated by the conventional letters:—

$l=0$	1	2	3	4	5
s	p	d	f	g	h

and the value of q is written as an index, thus:—$(nl)^q$. A configuration is usually specified by the $(nl)^q$ symbols indicating the occupied wave functions, as for example,

$$Cu^+ \quad (1s)^2(2s)^2(2p)^6(3s)^2(3p)^6(3d)^9(4s);$$

for brevity, the complete groups are often omitted as understood, so that this configuration would be written

$$Cu^+ \quad (3d)^9(4s).$$

For a configuration of closed groups and a few outer wave functions, the closed groups are often referred to collectively as the "core". The core charge will be written C. If some of the inner wave functions are unoccupied, as for the states of high excitation which give the terms of the x-ray spectra, it is more convenient to indicate the (nl) values of the *un*occupied wave functions (see White 1934, pp. 309, 310); for example,

$$Cu^+ \quad (1s)^{-1}$$

would represent a copper atom "ionized in the K shell" in the more usual terminology.

A configuration†, consisting of more than one wave function in addition to complete groups, gives not a single level but a number of levels† which, in Russell-Saunders coupling, can be grouped into terms† each specified by a total orbital moment of momentum L, and spin moment of momentum S, and, to the approximation contemplated, the wave function for any one of these terms is not a single determinant (1), but a linear combination

$$\Psi = \Sigma_A C_A \Psi(A; R) \qquad \ldots\ldots(3)$$

of such determinants, the sets A of wave functions in the different determinants including different members of the sets of one-electron wave functions available to the electrons not in the complete groups (see Slater (1929), Condon and Shortley (1935), Ch. VIII, and Kramers (1938)).

§ 2. THE VARIATION PRINCIPLE

The approximation expressed by the use of (1) and (2) as expressions for the wave functions of a many-electron atom determines everything except the radial wave functions of the different (nl) groups of the configuration considered. Each of these is a function of one variable only, so that we have only to determine a set of functions, in number equal to the number of (nl) groups in the configuration and each of one variable, to specify the whole atomic structure in this approximation. What is needed still is a means of obtaining the appropriate radial wave functions. Since the use of (1) and (2) is essentially an approximation, there is no such thing as an "exact" set of such functions; all we can expect is to find a "best" set according to some criterion.

A method both for finding approximate wave functions, and, for the normal state at least, of judging between different approximations is provided by the

† The words "level", "term" and "configuration" are used here in the senses defined by Condon and Shortley (1935), pp. xiii, 168, 189.

variation principle, which states that if for a system whose Hamiltonian function is H, we consider the quantity

$$E = \int \Psi^* H \Psi \, d\tau \Big/ \int \Psi^* \Psi \, d\tau \qquad \ldots\ldots(4)$$

evaluated for all functions Ψ for which $\int \Psi^* \Psi \, d\tau$ exists, then the functions Ψ for which E has stationary values are the wave functions of the stationary states. The stationary values of E are the energies of the system in the corresponding states.

Further, for the normal state the stationary value of the quantity E defined by (4) is an absolute minimum, so that if one function Ψ gives a smaller value of E than another, the smaller must be nearer the value which would be obtained by using the actual wave function for the normal state, supposing it is known. This gives a criterion for comparing different approximate wave functions; one such function can be regarded as a "better" approximation to the normal state if it gives a smaller value of E. It should be noted that "better" in this sense means "better for the calculation of the total energy of the whole atomic system"; other approximations might be better for calculating other atomic properties, but in most cases the only test is the empirical one of agreement with the observed values. No one criterion can give a full statement of the goodness of an approximate wave function, since for different applications the relative importance of different features, such as behaviour for small r, for large r, in the neighbourhood of the maximum of Ψ etc., will be different. But the energy criterion seems at present to be the most generally useful one. Other possible criteria for approximate wave functions have been discussed and compared with the energy criterion by James and Coolidge (1937) and by Williamson (1942). The appropriate wave functions to take for hyperfine structure calculations have been considered by Shoupp, Bartlett and Dunn (1935), Wills and Breit (1935) and Gibbons and Bartlett (1945).

For an excited state, E defined by (4) is an absolute minimum for all functions Ψ which are orthogonal to those of all lower states; but the wave functions of the lower states will only be known approximately, so this condition cannot be applied in an exact form. There are two possible alternative procedures, either to omit, or to attempt to apply, a condition of orthogonality, and in principle neither is entirely satisfactory. Suppose no such condition is applied; then if only a crude form of trial wave function for an excited state is taken, the result of applying the variation principle may only be a bad approximation to the normal state instead of an approximation to the excited state. On the other hand, insistence on orthogonality to what is only an approximation to the wave function for the normal state may lead to a worse approximation to that for the excited state than would be obtained without this condition. In practice, it seems possible to obtain useful approximations to wave functions for excited states without applying a formal condition of orthogonality to the approximate lower states.

In determining radial wave functions, it is usual to take for the Hamiltonian function H that for a system of point electrons, namely

$$H = \Sigma_i \left[-\tfrac{1}{2}\nabla_i^2 - \frac{N}{r_i} \right] + \Sigma'_{ij} \frac{1}{r_{ij}}, \qquad \ldots\ldots(5)$$

Σ' here indicates a sum over all pairs of values of (ij) excluding $i=j$, each pair being counted once.

There are two ways of using the variation principle to find approximate wave functions of atomic systems. The first is to consider functions Ψ of specified analytical form containing arbitrary *parameters* for which values have to be found so as to satisfy the variation principle for functions of this specified form. For example, we might restrict the radial wave functions to be considered to products of exponentials and polynomials, and find the "best" values of coefficients in the exponents and in the polynomials. This process will give not the actual wave function, except in very special cases, but the "best" approximation to it, in the sense already specified, obtainable with functions of the form specified.

This kind of approximation is considered in § 7. A general survey of the application of the variation principle to the determination of approximate wave functions of atoms and molecules, with particular reference to analytic wave functions, has been given by James (1941).

The second way is to try a wave function constructed in a specified manner out of arbitrary *functions* of the variables of the problem, and determine these functions so as to satisfy the variation principle. Again, we may know that we cannot get an exact solution of the form adopted, but may still ask what is the "best" solution of this form we can get; and the additional freedom, compared with the use of analytical functions of specified form, means that it should be possible to get a better approximation. The equations for the radial wave functions obtained by substituting approximate wave functions given by (1) or (3), as appropriate, and (2) into the variation principle were first derived by Fock (1930). Other derivations and discussions have been given by Dirac (1931), Lennard-Jones (1933), and Shortley (1936). The first solution was evaluated by Fock and Petrashen (1934). This type of approximation is considered in §§ 3-6.

There are two steps in the derivation of these equations for any particular atomic state; first the derivation of the formula for the energy E in terms of integrals over the radial wave functions, and secondly the application of the variation principle to this expression for E so as to obtain the equations for the "best" radial wave functions.

§ 3. EXPRESSION FOR THE TOTAL ENERGY IN TERMS OF THE RADIAL WAVE FUNCTIONS

(a) Ψ *a single determinant of one-electron wave functions*

When the appropriate wave function for the whole atomic system is a *single* determinant (1), a convenient way of deriving the formula for E is by use of the set of coefficients defined as follows :—

$\epsilon_{\alpha\beta\gamma\ldots\pi} = 0$ if $\alpha, \beta, \gamma, \ldots \pi$ are not all unequal,

$= 1$ if $\alpha, \beta, \gamma, \ldots \pi$ form an even permutation of the natural numbers $1, 2, 3, \ldots p$,

$= -1$ if $\alpha, \beta, \gamma, \ldots \pi$ form an odd permutation of the natural numbers $1, 2, 3, \ldots p$.

If the label of a wave function is temporarily written as a suffix, thus:— $\psi_\alpha(j)$, $\psi_\beta(k)$, ..., and the summation convention is understood, the determinant (1) can be written

$$\Psi = \epsilon_{\alpha\beta\gamma\ldots\pi}\psi_\alpha(1)\psi_\beta(2)\ldots\psi_\pi(p).$$

The manipulation of integrals of the form $\int \Psi^*[\Sigma_i f(i)]\Psi\, d\tau$ and $\int \Psi^*[\Sigma'_{ij} g(ij)]\Psi\, d\tau$ in terms of this notation for the determinant (1) has been given by Waller and Hartree (1929), and can be applied directly to the integrals occurring in the formula (3) for the energy with the Hamiltonian (5).

The coefficients $\epsilon_{\alpha\beta\gamma\ldots\pi}$ have the properties:—

$$\epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha\beta\gamma\ldots\pi} = p! \qquad \ldots\ldots(6)$$

$$\epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta\gamma\ldots\pi} = (p-1)!\,\delta_{\alpha\alpha'} \qquad \ldots\ldots(7)$$

$$\epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta'\gamma\ldots\pi} = (p-2)!\,(\delta_{\alpha\alpha'}\delta_{\beta\beta'} - \delta_{\alpha\beta'}\delta_{\beta\alpha'}) \qquad \ldots\ldots(8)$$

(Waller and Hartree 1929), the summation convention being again understood. As already mentioned, it is no restriction to assume the one-electron wave functions to be orthogonal; they will also be supposed normalized, so that

$$\int \psi^*_\alpha(i)\psi_\beta(i)\, d\tau_i = \delta_{\alpha\beta}. \qquad \ldots\ldots(9)$$

With the Hamiltonian (5), the expression for E involves three kinds of integrals, namely

$$\int \Psi^*\Psi\, d\tau, \qquad \int \Psi^*\Sigma_i[-\tfrac{1}{2}\nabla_i^2 - N/r_i]\Psi\, d\tau, \qquad \int \Psi^*\Sigma'_{ij}(1/r_{ij})\, d\tau.$$

The results can be expressed in terms of integrals involving single members and pairs of the set of one-electron wave functions ψ; it is convenient to use the following notation, introduced by Slater (1929), for these integrals:—

$$\mathrm{I}_\alpha = -\tfrac{1}{2}\int \psi^*_\alpha(i)[\nabla_i^2 + 2N/r_i]\psi_\alpha(i)\, d\tau_i, \qquad \ldots\ldots(10)$$

$$\mathrm{J}_{\alpha\beta} = \int\int |\psi_\alpha(i)|^2(1/r_{ij})|\psi_\beta(j)|^2\, d\tau_i\, d\tau_j, \qquad \ldots\ldots(11)$$

$$\mathrm{K}_{\alpha\beta} = \int\int \psi^*_\alpha(i)\psi_\beta(i)(1/r_{ij})\psi_\alpha(j)\psi^*_\beta(j)\, d\tau_i\, d\tau_j. \qquad \ldots\ldots(12)$$

Consider first $\int \Psi^*\Psi\, d\tau$. This is

$$\epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta'\gamma'\ldots\pi'} \int \psi^*_\alpha(1)\psi_{\alpha'}(1)\, d\tau_1 \int \psi^*_\beta(2)\psi_{\beta'}(2)\, d\tau_2 \ldots \int \psi^*_\pi(p)\psi_{\pi'}(p)\, d\tau_p$$

$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta'\gamma'\ldots\pi'}\delta_{\alpha\alpha'}\,\delta_{\beta\beta'}\ldots\delta_{\pi\pi'} \qquad \text{[from (9)]}$$

$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\epsilon_{\alpha\beta\gamma\ldots\pi} = p! \qquad \text{[from (6)].}$$

In considering the sum

$$\int \Psi^*\Sigma_i[-\tfrac{1}{2}\nabla_i^2 - N/r_i]\Psi\, d\tau,$$

write, for brevity, $f(i)$ for $[-\tfrac{1}{2}\nabla_i^2 - N/r_i]$ and $f_{\alpha\alpha'}$ for $\int \psi^*_\alpha(i)[-\tfrac{1}{2}\nabla_i^2 - N/r_i]\psi_{\alpha'}(i)d\tau_i$.

Take *one* term of the sum $\int \Psi^*\Sigma_i f(i)\Psi\, d\tau$, say that with $i=1$. This is

$$\int \Psi^* f(1)\Psi\, d\tau = \epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta'\gamma'\ldots\pi'} \int \psi^*_\alpha(1)f(1)\psi_{\alpha'}(1)\, d\tau_1 \int \psi^*_\beta(2)\psi_{\beta'}(2)\, d\tau_2$$

$$\ldots. \int \psi^*_\pi(p)\psi_{\pi'}(p)\, d\tau_p$$

$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta'\gamma'\ldots\pi'} f_{\alpha\alpha'}\delta_{\beta\beta'}\ldots\delta_{\pi\pi'} \qquad \text{[from (9)]}$$

$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\,\epsilon_{\alpha'\beta\gamma\ldots\pi}\, f_{\alpha\alpha'} = (p-1)!\,\delta_{\alpha\alpha'} f_{\alpha\alpha'} \qquad \text{[from (7)]}$$

$$= (p-1)!\, f_{\alpha\alpha} = (p-1)!\,\Sigma_\alpha \mathrm{I}_\alpha.$$

There are p such terms in $\int \Psi^* \Sigma_i [-\frac{1}{2}\nabla_i^2 - N/r_i]\Psi \, d\tau$, so that altogether

$$\int \Psi^* \Sigma_i [-\tfrac{1}{2}\nabla_i^2 - N/r_i]\Psi \, d\tau = p\,! \, \Sigma_\alpha I_\alpha ;$$

it will be noted that a sum over all *electrons* on the left-hand side has been transformed into a sum over all the *one-electron wave functions* on the right.

Similarly, consider *one* of the terms of the sum $\int \Psi^* \Sigma'_{ij} (1/r_{ij}) \Psi \, d\tau$, say that with $i=1, j=2$; for brevity, write

$$u_{\alpha\alpha',\beta\beta'} \text{ for } \int\int \psi_\alpha^*(1)\psi_\beta^*(2)(1/r_{12})\psi_{\alpha'}(1)\psi_{\beta'}(2)\, d\tau_1 \, d\tau_2 .$$

Then this contribution to the sum is

$$\int \Psi^*(1/r_{12})\Psi \, d\tau = \epsilon_{\alpha\beta\gamma\ldots\pi}\, \epsilon_{\alpha'\beta'\gamma'\ldots\pi'} \int\int \psi_\alpha^*(1)\psi_\beta^*(2)(1/r_{12})\psi_{\alpha'}(1)\psi_{\beta'}(2)\, d\tau_1 \, d\tau_2$$
$$\int \psi_\gamma^*(3)\psi_{\gamma'}(3)\, d\tau_3 \ldots \int \psi_\pi^*(p)\psi_{\pi'}(p)\, d\tau_p$$
$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\, \epsilon_{\alpha'\beta'\gamma'\ldots\pi'}\, u_{\alpha\alpha',\beta\beta'}\, \delta_{\gamma\gamma'} \ldots \delta_{\pi\pi'} \qquad \text{[from (9)]}$$
$$= \epsilon_{\alpha\beta\gamma\ldots\pi}\, \epsilon_{\alpha'\beta'\gamma\ldots\pi}\, u_{\alpha\alpha',\beta\beta'} = (p-2)\,!\,(\delta_{\alpha\alpha'}\delta_{\beta\beta'} - \delta_{\alpha\beta'}\delta_{\beta\alpha'})u_{\alpha\alpha',\beta\beta'} \qquad \text{[from (8)]}$$
$$= (p-2)\,!\,(u_{\alpha\alpha,\beta\beta} - u_{\alpha\beta,\beta\alpha}) = (p-2)\Sigma_{\alpha\beta}(J_{\alpha\beta} - K_{\alpha\beta}).$$

There are $\frac{1}{2}p(p-1)$ terms $(1/r_{ij})$ in the integrand, each of which gives a contribution of this amount, so that altogether

$$\int \Psi^* \Sigma_{ij} (1/r_{ij}) \Psi \, d\tau = \tfrac{1}{2}(p\,!)\Sigma_{\alpha\beta}(J_{\alpha\beta} - K_{\alpha\beta}).$$

The sum on the right is over *all* values of α and *all* values of β separately; any particular pair of values appears twice. If $\alpha = \beta$, $J_{\alpha\beta} = K_{\alpha\beta}$, so these integrals can be either included in both sums or omitted in both; it is more convenient to omit them. If the sum over all pairs α, β, excluding $\alpha = \beta$ and counting each pair *once*, is written $\Sigma'_{\alpha\beta}$, then

$$\int \Psi^* \Sigma_{ij} (1/r_{ij}) \Psi \, d\tau = p\,! \, \Sigma'_{\alpha\beta}(J_{\alpha\beta} - K_{\alpha\beta}).$$

From the orthogonality of the spin wave functions, the $K_{\alpha\beta}$ integrals are zero unless ψ_α and ψ_β have the same spin; it is convenient to indicate by $\Sigma''_{\alpha\beta}$ the sum taken over all pairs of wave functions with the same spin, excluding $\alpha = \beta$. So finally the expression for the energy is

$$E = \Sigma_\alpha I_\alpha + \Sigma'_{\alpha\beta} J_{\alpha\beta} - \Sigma''_{\alpha\beta} K_{\alpha\beta} \qquad \ldots\ldots (13)$$

(see Slater 1929, p. 1307).

The argument so far depends on the wave functions for the whole atom being expressed as a *single* determinant (1) of the one-electron wave functions ψ, but does not depend on these wave functions being of central-field form. The single determinant wave function (1) for a whole atom is appropriate to configurations which give each a single term; that is, to configurations of closed groups and one other wave function, and configurations of closed groups except for the lack of one wave function; the extension of the results to other configurations will be considered later.

If for an atom, the further restriction to central-field wave functions is made,

the integrations with respect to the spherical polar angles can be carried out and the integrals (10), (12) reduced to integrals over the radial coordinates only, and finally gives a formula for E in terms of integrals of the following three types:—

$$\mathrm{I}(nl) = -\tfrac{1}{2}\int_0^\infty \mathrm{P}(nl;r)\left[\frac{d^2}{dr^2} + \frac{2N}{r} - \frac{l(l+1)}{r^2}\right]\mathrm{P}(nl;r)\,dr, \qquad \ldots\ldots(14)$$

$$\mathrm{F}_k(nl, n'l') = \int_0^\infty\int_0^\infty [\mathrm{P}(nl;r_1)]^2[\mathrm{P}(n'l';r_2)]^2(r_a{}^k/r_b{}^{k+1})\,dr_1\,dr_2, \qquad \ldots\ldots(15)$$

$$\mathrm{G}_k(nl, n'l') = \int_0^\infty\int_0^\infty \mathrm{P}(nl;r_1)\mathrm{P}(n'l';r_1)\mathrm{P}(nl;r_2)\mathrm{P}(n'l';r_2)(r_a{}^k/r_b{}^{k+1})\,dr_1\,dr_2, \qquad \ldots\ldots(16)$$

where r_a is the smaller and r_b the larger of r_1 and r_2; note that $\mathrm{G}_k(nl, nl)$ and $\mathrm{F}_k(nl, nl)$ are the same. The expression for E is of the general form

$$E = \Sigma_{nl} q(nl)\mathrm{I}_{nl} + \Sigma_{nl,\,n'l',\,k} \mathrm{A}_{ll'k}\mathrm{F}_k(nl, n'l') - \Sigma_{nl,\,n'l',\,k}\mathrm{B}_{ll'k}\mathrm{G}_k(nl, n'l'). \qquad \ldots\ldots(17)$$

The coefficient $q(nl)$ of $\mathrm{I}(nl)$ is the number of occupied wave functions in the (nl) group, and is $2(2l+1)$ for a complete group. The coefficients $\mathrm{A}_{ll'k}$ and $\mathrm{B}_{ll'k}$ for the integrals referring to complete (nl) and $(n'l')$ groups have been tabulated by D. R. and W. Hartree (1936).

For a configuration consisting of complete groups and *one* additional wave function, let (nl) refer to the additional wave function and $(n'l')$ to the core; then $q(nl) = 1$ and the coefficients $\mathrm{A}_{ll'k}$ and $\mathrm{B}_{ll'k}$ are $1/2(2l+1)$ times those for a complete (nl) group (D. R. and W. Hartree, 1937, p. 168).

(b) *Ψ a linear combination of determinants of one-electron wave functions*

When there are more wave functions than one, in addition to closed groups, a single configuration gives rise to more than one term, and the appropriate wave function Ψ for each of these terms is generally not a single determinant (1) but a linear combination (3) of such determinants. However, for many of the simpler configurations it is possible to derive expressions for the *energies* of the different terms without explicitly determining the appropriate linear combinations of determinants (see Slater, 1929, p. 1304; Condon and Shortley, 1935, Chap. 7, § 2; van der Waerden, 1932, §§ 28, 29). The expressions for the energy are still of the general types (14), (15), (16), but with different coefficients $\mathrm{A}_{ll'k}$ and $\mathrm{B}_{ll'k}$.

The contributions to E can be divided into three groups, namely:—

(i) those arising from the core of complete groups only;

(ii) those arising from the I integrals for wave functions outside the core of closed groups, and the F_k and G_k integrals for the interactions between the electrons in each of these wave functions and the core;

(iii) the F_k and G_k integrals arising from the interactions between the electrons in the outer wave functions.

The differences between the energy formulae for the different terms arising from a single configuration is restricted to the contributions of group (iii).

The perturbation of a core of complete groups on adding electrons in wave functions lying mainly outside the core is usually small, particularly in the case of added wave functions with values of n greater than those of the core wave functions,

and can often be neglected. Then it is convenient to reckon energies from that of the core alone as zero, and both in the derivation of the equations of the outer wave functions and of the energy values relative to that zero, the contributions of group (i) to the energy can be omitted.

§ 4. DERIVATION OF EQUATIONS FOR THE RADIAL WAVE FUNCTIONS

The expression for the energy in terms of I, F_k and G_k integrals is derived on the understanding that the one-electron wave functions (2) are normalized and orthogonal. This condition is satisfied provided the radial wave functions satisfy

$$\int_0^\infty \mathrm{P}(nl; r)\mathrm{P}(n'l'; r)\,dr = \delta_{nn'}, \qquad \ldots\ldots(18)$$

and in applying the variation principle to obtain the equations for the "best" radial wave functions $\mathrm{P}(nl; r)$, the variations must be made subject to these conditions. This can be ensured by introducing Lagrange multipliers $\lambda_{nl,\,n'l}$ and considering the variation not of E but of the function

$$E' = E + \Sigma_{nn'}\lambda_{nl,\,n'l}\int_0^\infty \mathrm{P}(nl\,; r)\mathrm{P}(n'l\,; r)\,dr,$$

with respect to the $\mathrm{P}(nl; r)$'s regarded as independent, the quantities $\lambda_{nl,\,n'l}$ being adjustable, to be determined later so that the conditions (18) are satisfied.

The equations for the functions $\mathrm{P}(nl; r)$ can be derived by formal differentiation with respect to each of these functions (see D. R. and W. Hartree (1936), p. 591). In expressing the results it is convenient to write $\mathrm{Y}_k(nl, n'l'; r)$ for the function

$$\mathrm{Y}_k(nl, n'l'; r) = \int_0^r \mathrm{P}(nl; r_1)\mathrm{P}(n'l'; r_1)(r_1/r)^k\,dr_1 + \int_r^\infty \mathrm{P}(nl; r_1)\mathrm{P}(n'l'; r_1)(r/r_1)^{k+1}\,dr_1 \qquad \ldots\ldots(19)$$

$r^{-1}\mathrm{Y}_0(nl, n'l'; r)$ is the potential at r of a spherically symmetrical charge distribution of radial density $\mathrm{P}(nl; r)\mathrm{P}(n'l'; r)$. If $\mathrm{S}_k(\theta, \phi)$ is a spherical harmonic of order k, $\mathrm{Y}_k(nl, n'l'; r)\mathrm{S}_k(\theta, \phi)$ is similarly related to the potential of a charge distribution with charge density $[\mathrm{P}(nl; r)\mathrm{P}(n'l'; r)/r^2]\mathrm{S}_k(\theta, \phi)$. It is further convenient to write $\mathrm{T}(r)$ for the quantity

$$\mathrm{T}(r) = N - \Sigma_{n'l'}q(n'l')\mathrm{Y}_0(n'l'; r). \qquad \ldots\ldots(20)$$

If one electron is thought of as occupying each of the wave functions $\psi(\alpha; j)$, then $\mathrm{T}(r)/r$ is the spherical average of the potential at radius r of the Schrödinger charge distribution of the electrons.

Differentiation of E' with respect to each $\mathrm{P}(nl; r)$ then gives a set of equations of the general form

$$\Sigma_{n'\Gamma}H'_{nl,n'\Gamma}\mathrm{P}(n'l'; r) = 0. \qquad \ldots\ldots(21)$$

For a configuration of closed groups, the diagonal elements of the matrix H' are

$$H'_{nl,\,nl} = \frac{d^2}{dr^2} + \frac{2}{r}\left[\mathrm{T}(r) + \mathrm{Y}_0(nl, nl; r) + \frac{1}{2l+1}\Sigma_k \mathrm{A}_{jk}\mathrm{Y}_k(nl, nl; r) - \epsilon_{nl,\,nl} - l(l+1)/r^2\right] \qquad \ldots\ldots(22)$$

and the non-diagonal elements are

$$H'_{nl,\,n'l'}=\Sigma_k \frac{1}{2l+1} B_{ll'k} \frac{\gamma_k(nl,\,n'l')}{r} - \epsilon_{nl,n'l'}\delta_{ll'}. \quad \ldots\ldots(23)$$

These equations were first derived by Fock (1930) and are usually known as Fock's equations. For other configurations, the expressions are of the same general form but with other coefficients in place of A_{llk} and B'_{llk} (see Shortley 1936). The $\epsilon_{nl,\,n'l}$ parameters in (22), (23) are simply related to the Lagrange multipliers $\lambda_{nl,\,n'l}$ in the definition of E'.

Some examples of Fock's equations, written out in full for particular cases, will be found in Fock (1930), Brown (1933), Hartree, Hartree and Manning (1941 c).

If the terms with $k>0$ in (22), and the non-diagonal elements (23) are omitted, equation (21) simplifies to

$$\left[\frac{d^2}{dr^2}+\frac{2}{r}\{T(r)+Y_0(nl,\,nl;\,r)\}-\epsilon_{nl,\,nl}-\frac{l(l+1)}{r^2}\right]P(nl;\,r)=0. \quad \ldots\ldots(24)$$

This set of equations was derived earlier by Hartree (1929) without appeal to the variation principle, in an attempt to express in quantitative form, in terms of wave mechanics, the approximation in which each electron is thought of as being in a stationary state in the field of the nucleus and of the space-charge distribution of the other electrons. The method of evaluating solutions of these equations, by estimating contributions to the field from the different (nl) groups and adjusting these estimates until the field derived from the space-charge distributions calculated from the solutions $P(nl;\,r)$ reproduces the estimated contributions, led to the use of the term "self-consistent field" for the field so calculated.

In this treatment, Pauli's principle is taken into account only in so far as it restricts the number of wave functions in each (nl) group, and a further approximation is made in taking a spherical average of the electron distributions in wave functions with $l>0$ in evaluating their contributions to the field acting on other electrons. An analytical study of this approximation was made by Gaunt (1928) and by Slater (1928), and its relation to the variation principle was pointed out by Slater (1930) and Fock (1930).

In the derivation of Fock's equations for a configuration of closed groups, Pauli's principle is taken into account by the use of the approximate form (1) for the wave function. This can be described verbally as expressing the fact that no one electron can be assigned to a particular one of the one-electron wave functions $P(nl;\,r)$, but that, if this kind of approximation is adopted at all, there must be "exchange" of the different electrons between these wave functions. The use of this determinant for Ψ, instead of a simple product, gives rise to the non-diagonal elements (23) ,which give in equations (21) terms which are usually called "exchange terms". The terms with $k>0$ in (22) arise partly from similar terms expressing exchange within a single (nl) group, and partly from departures from spherical symmetry of the wave functions for $l>0$. These two contributions are of the same order of magnitude, and there is no point in including the latter only (Slater 1930). Brown (1933) and Kennard and Ramberg (1934) have used a form of "self-consistent field" in which these exchange terms and terms arising from non-spherical symmetry within each complete group are included, but the non-diagonal elements $H_{nl,\,n'l'}$ are neglected.

Since the main and characteristic difference by which equations (21) differ from (24) is in the appearance of the exchange terms in the former, the field derived from the solution of Fock's equation (21) is often called "the self-consistent field with exchange".

In applying the method of the self-consistent field (either without or with exchange), to excited states, it has not been found necessary to make the wave function for an excited state orthogonal to those for all lower states. It appears possible to obtain as good an approximation to an excited state as to the normal state without imposing this condition, which would involve a considerable amount of extra work in evaluating the solutions. It should be noted, though, that Mosse, Young and Haurwitz (1935) found that if wave functions were restricted to certain specified analytical forms which would be expected to give a good approximation, no stationary value of E corresponding to the $(1s)\,(2s)\,{}^1S$ state of He could be found. Wilson and Lindsay (1935), however, found no difficulty in obtaining a "self-consistent field" solution for this state.

Calculation of Energies.—It is often required to calculate energy values from radial wave functions. On account of the stationary property of the energy, evaluated from (3), for variations in the wave functions, it is valid to include exchange terms in the evaluation of the total energy although such terms have been omitted in calculation of the radial wave functions themselves, but the calculation is simpler, and the energy values better, if the wave functions have been calculated from the solution of Fock's equations.

Sometimes the total energy of the whole atomic system may be required, but more often it is the energy relative to some higher state of ionization, usually a configuration of closed groups, taken as zero; for example, for Al $(3s)^2(3p)$, the the energy relative to the normal state of Al^{+3} might be required. The wave functions of the core are very little affected by the removal of electrons in outer wave functions, and on account of the stationary property of the energy of the core for variation in the core wave functions, the variation of the contribution (i) of § 3(b) to the energy can usually be neglected. For the contribution (ii) it is most convenient to carry out algebraically the substitution for

$$\left[\frac{d^2}{dr^2}+\frac{2N}{r}-\frac{l(l+1)}{r^2}\right]\mathrm{P}(nl;r)$$

in the $\mathrm{I}(nl)$ integrals in terms of the differential equation satisfied by $\mathrm{P}(nl;r)$; this reduces the value of each $\mathrm{I}(nl)$ to linear combinations of F_k and G_k integrals, which would have to be evaluated in any case, and $\epsilon_{nl,\,nl}$ (for examples, see D. R. and W. Hartree 1936a, 1938a, and Hartree, Hartree and Swirles 1939).

These expressions are simpler and easier to evaluate if the radial wave functions are obtained from Fock's equations than if they are obtained from the equations of the self-consistent field without exchange. For the particular case of one electron outside closed groups, the energy $E(nl)$ relative to the energy of the core as zero is just given by

$$2E(nl)=-\epsilon_{nl,\,nl}$$

(D. R. and W. Hartree 1936a, § 7, see also Fock 1933) if $\mathrm{P}(nl;r)$ is calculated from Fock's equation, so that no further calculation is required, whereas if exchange terms are neglected in the calculation of $\mathrm{P}(nl;r)$, several more steps of calculation are required to obtain $E(nl)$ (McDougall 1932).

§ 5. THE PROCESS OF SOLVING THE EQUATIONS

In any problem of the "self-consistent field" type there are two kinds of relations to be satisfied, the equations of motion of the particles in the field in classical mechanics, or the corresponding wave equation in quantum mechanics, and the equations (Poisson's equation or equations derived from it) for the field due to the charge distribution of the particles so moving. In the present state these equations (21)–(23) or (24) and the relations (19) respectively.

Although (19) gives the functions Y_k in terms of quadratures, this is usually not the most convenient form from which to calculate them (McDougall 1932, p. 568, and Hartree and Black 1933, p. 326; Yost (1940), however, regards the evaluation of the Y_k's by double quadrature as quite practicable). If

$$Z_k(nl, n'l'; r) = \int_0^r P(nl; r_1)P(n'l'; r_1)(r_1/r)^k dr_1, \qquad \ldots\ldots(25)$$

then (McDougall 1932, p. 568)

$$\frac{dZ_k}{dr} = P(nl; r)P(n'l'; r) - \frac{k}{r}Z_k, \qquad \ldots\ldots(26)$$

and

$$\frac{dY_k}{dr} = \frac{1}{r}[(k+1)Y_k - (2k+1)Z_k]; \qquad \ldots\ldots(27)$$

Z_k and Y_k are solutions of these equations with boundary conditions

$$\left.\begin{aligned} Z_k(nl, n'l'; 0) &= 0 \\ \operatorname*{Lt}_{r\to\infty}[r^k Y_k(nl, n'l'; \infty)] &= \int_0^\infty r^k P(nl; r)P(n'l'; r)\,dr; \end{aligned}\right\} \qquad \ldots\ldots(28)$$

for two alternative forms of the equations, one possibly more convenient for small r and the other for large r (Hartree and Black 1933, § 7). The quantities Z_k/r^2 give the radial variations of contributions to the *field*, corresponding to the contributions Y_k/r to the *potential*. The most convenient procedure seems to be to integrate (26) outwards, as it stands or in a modified form, and then to integrate (27) inwards. It is possible to avoid use of the function Z_k altogether by taking equations (25), (26) in the form

$$\frac{d^2Y_k}{dr^2} = \frac{k(k+1)}{r^2}Y_k - \frac{2k+1}{r}P(nl; r)P(n'l'; r), \qquad \ldots\ldots(29)$$

but the solution of this equation has to satisfy one condition at each end of the range of integration, and this is a situation to be avoided if it can be avoided easily, as is the case by the use of (26), (27).

The evaluation of solutions of equations (21)–(23) or (24), and (26), (27) has usually been carried out by some form of successive approximation. Let $P_{(0)}(nl; r)$ be a set of estimates of the radial wave functions $P(nl; r)$. Then from the functions $P_{(0)}$ regarded as given, corresponding functions Z_k can be calculated by the solution of equation (26), then corresponding functions Y_k can be calculated by the solution of equations (27), and finally solutions of equations (21)–(23) or (24) with these functions Y_k can be evaluated; let these "calculated" or "final" solutions be written $P_{(f)}(nl; r)$. If the "initial" or "estimated" functions have been taken such that

$$P_{(f)}(nl; r) = P_{(0)}(nl; r), \qquad \ldots\ldots(30)$$

they form a solution of the whole system of equations. The method is then to make estimates $P_{(i)}(nl; r)$ and to adjust them until the condition (30) is satisfied to an adequate approximation.

One possible way of adjusting the estimates is to make the method iterative by taking the estimates $P_{(i)}(nl; r)$ for one stage of the calculation as the final functions $P_{(f)}(nl; r)$ of the previous stage. For highly ionized atoms, and the inner wave-functions of any atom, this process is satisfactory, but for the outer wave-functions of slightly-ionized, neutral and negatively-ionized atoms it may only converge slowly, or may even diverge (Hartree 1933), and it is better, and may be necessary, to adopt a less automatic method of adjusting the estimates; experience of similar calculations for other atoms and ions is then the best guide.

In the approximation of the self-consistent field without exchange, expressed by equation (24), the only functions $Y_k(nl, n'l'; r)$ occurring are those with $k=0$ and $(n'l')=(nl)$, that is, functions $Y_0(nl, nl; r)$, and the number of them is the same as that of the radial wave functions. In this case, the usual procedure has been to start from estimates of the functions $Z_0(nl, nl; r)$, or rather, in practice, of $2q(nl)[1-Z_0(nl, nl; r)]$, then to derive corresponding contributions to $T(r)$, then to evaluate solutions of (24), and lastly to derive "final" functions $Z_0(nl, nl; r)$, and to adjust the estimates until the "final" functions agree with them. But when exchange terms are included, the number of Z_k functions is considerably larger than the number of radial wave functions (for example, for Cu^+ there are six radial wave functions and 28 Z_k functions, whereas for Rb^+ the corresponding numbers are eight and 50 (see D. R. and W. Hartree (1935), p. 491) so that it is better to start from estimates of the radial wave functions.

The first estimates can be made in various ways, depending on what calculations have been carried out for neighbouring atoms in the periodic table. Interpolation with respect to the atomic number N can often be carried out satisfactorily by a scaling procedure, in which each radial wave function is taken to be of the form

$$P(nl; r)=(N-s_{nl})^{\frac{1}{2}}\,Q_{nl}\,[(N-s_{nl})r],$$

where s_{nl} is a screening number for the (nl) wave function, and $Q_{nl}(z)$ a function which for fixed z varies slowly with N. An example of reduced radial wave functions obtained by plotting $P(nl; r)/(N-s_{nl})^{\frac{1}{2}}$ against $(N-s_{nl})r$ is given in figure 1, which refers to $(2p)$ wave functions calculated with exchange. If the functions $Q_{nl}(z)$ were independent of N, then the functions $Z_0(nl, nl; r)$ for different atoms would be functions of $(N-s_{nl})r$ only, so that the graphs of $Z_0(nl, nl; r)$ against $\log r$ would be replicas of the same curve shifted horizontally from one another. Actually such a relation is only approximate, but provides a good basis for initial estimates of the $Z_0(nl, nl; r)$ functions in calculations of the self-consistent field without exchange (Arnot and McLauchlan 1934, Manning and Millman 1936).

Another method of making the estimates is based on interpolating not the wave functions $P(nl; r)$, but their departures from scaled hydrogen wave functions (Hartree 1929).

Since equations (24) for the self-consistent field without exchange are homogeneous, solutions can be evaluated without particular attention to normalization, and each normalized afterwards by division by $\left[\int_0^\infty \{P(nl; r)\}^2\,dr\right]^{\frac{1}{2}}$. But if the

functions Z_k are given, equations (21)–(23) for the self-consistent field with exchange are inhomogeneous, and attention must be then paid to normalization in the course of the solution. Experience has shown that it is best to aim at working with normalized wave functions throughout, by taking the quantities $[P(nl; r)/r^{l+1}]_{r=0}$ as parameters to be chosen so that the solutions of equations (21)–(23) are normalized, rather than using unnormalized wave functions and explicit normalizing factors, as was done in some early work.

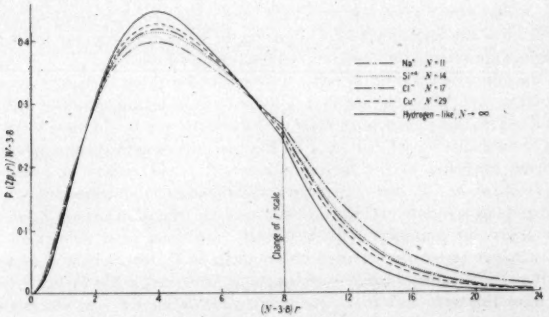


Figure 1. Example of scaling of radial wave functions $P(2\rho; r)/(N-3{\cdot}8)^{\frac{1}{2}}$ plotted against $(N-3{\cdot}8)r$. Radial wave functions $P(2\rho; r)$ calculated with exchange, and normalized.

Another way of carrying out the solution is by integrating the wave equations (21)–(23) or (24) and Poisson's equation (27) outwards *simultaneously*, instead of *alternately*. For the self-consistent field without exchange, for which the only Y_k functions are the functions $Y_0(nl, nl; r)$; this process has been used successfully by Torrance (1934) who adopted the following method of integrating the equations for $Y_0(nl, nl; r)$ simultaneously with those for $P(nl; r)$. From (26), (27) it follows that for $k=0$,

$$Y_0(nl, nl; r)/r = A_0 - \int_0 (1/r^2)\left\{\int_0 [P(nl; r)]^2\, dr\right\} dr, \quad \ldots\ldots(31)$$

where

$$A_0 = \int_0^\infty (1/r^2)\left[\int_0 \{P(nl; r)\}^2 dr\right] dr. \quad \ldots\ldots(32)$$

The two-fold integral in (31) can be evaluated by quadrature as the integration progresses, and though A_0 is not known until the solution is completed, it can be combined with $\epsilon_{nl, nl}$ which in any case has to be adjusted so as to make

$$P(nl; r) \to 0 \qquad \text{as} \qquad r \to \infty.$$

For $k \neq 0$, the relation corresponding to (31) is

$$\gamma_k(nl, n'l'; r)/r = (2k+1)r^k\left[A_k - \int_0 \{1/r^{2(k+1)}\}\left\{\int_0 r^k P(nl; r)P(n'l'; r)\, dr\right\} dr,\right] \quad \ldots\ldots(33)$$

where

$$A_k = \int_0^\infty \{1/r^{2(k+1)}\}\left\{\int_0 r^k P(nl; r)P(n'l'; r)\, dr\right\} dr. \quad \ldots\ldots(34)$$

Since $P(nl; r)$ is of order r^{l+1} at the origin, the quantity

$$\{1/r^{2(k+1)}\}\left\{\int_0 r^k P(nl; r) P(n'l'; r)\, dr\right\}, \qquad \ldots\ldots(35)$$

is of order $r^{l+l'-k+1}$; and since in the integrals $F_k(nl, n'l')$ and $G_k(nl, n'l')$ in the energy formula from which Fock's equations are derived, $k \leqslant (l+l')$, (35) is of order r^m where m is at least unity, and there is no difficulty in principle in the evaluation of the integral of (35). But for $k>0$ the value of A_k cannot be absorbed in other adjustable constants as the values of A_0 can, and if in the application of Torrance's method to Fock's equations, the Y_k's were calculated by outward evaluation of the integral in (33) concurrently with the solution of the equations for the radial wave functions, the A_k's in (33) would have to be regarded as parameters to be adjusted so that when the solution was completed the relation (34) would be satisfied. A similar use of adjustable parameters would have to be made if the Y_k's were obtained by numerical solution of equation (28) (D. R. and W. Hartree 1935, § 6). A comparison of such a method with a method of successive approximations, and alternate, rather than simultaneous, treatment of the equations of motion and of Poisson's equation has been made by Yost (1940), who comes to the conclusion that the latter is to be preferred as a practical method. He also reports that evaluation of the Y_k's by the repeated quadrature expressed by (33), rather than by solution of equations (26) and (27), is quite practicable.

The work involved in evaluating solutions of the equations of the self-consistent field has usually been carried out by purely numerical methods using a step-by-step method for the numerical integrations and a trial-and-error method for determining the characteristic values $\epsilon_{nl, nl}$; some details of a practical method for carrying out the numerical integrations have been given by Hartree (1933 a). In a few cases the differential analyser (Bush 1931) has been used for the integrations of the radial wave equation (Caldwell 1932, Hartree 1934, Porter 1935, Millard and Manning 1938). When this instrument is used, it seems most convenient to work with $\log r$ as independent variable and $Pr^{\frac{1}{2}}$ as dependent variable, to avoid changes of scale (Hartree 1934).

Another method of solving the equations, more allied to the "relaxation" methods developed by Southwell (1943, 1946) in other contexts, has been proposed by Kimball and Shortley (1934). It has not so far been used extensively for numerical work, but this or a similar method may be found to be most suitable in the application of large automatic digital calculating machines to this work, which is a possibility for the future.

§ 6. IMPROVEMENTS IN THE APPROXIMATION

(a) *Superposition of configurations*

The representation of an atomic state by a single configuration is of course an approximation; a very useful one, however, and, though perhaps somewhat crude, often capable of accounting qualitatively and approximately quantitatively for the main properties of the atomic state. Because of the relative simplicity of such an approximation, perhaps too much reliance has been placed on the use of a single configuration as an adequate approximation in most cases; but there are contexts in which a better approximation is desirable or even necessary. A general examination of the limitations of the approximation of representing

atomic states by configurations of one-electron wave functions has been made by Frank (1937).

There are two main ways of attempting to improve the approximation. One, which involves the rejection of the idea of "configurations" of one-electron wave functions, is considered in § 8. The other is to retain the idea of configurations but to use as an approximate wave function Ψ a linear combination of determinants (1) referring to different configurations.

The representation of an atomic state by a single configuration A is a significant approximation if the matrix components $(A|H|B)$ of the Hamiltonian for other configurations B are small compared with $(A|H|A)$. When this is not the case the appropriate wave function is a linear combination of Ψ_A and Ψ_B. This situation is usually expressed by the term "configuration interaction" (see, for example, Condon and Shortley 1935, Ch. 15); a better term is "superposition of configurations" (Hartree, Hartree and Swirles 1939, p. 230). Such superposition can only occur between wave functions of the same L, S, and parity (Condon and Shortley 1935, Ch. 15, § 1).

The equations of the self-consistent field can be extended to include superposition of configurations; the general equations for the case of the configurations $(1s)^2\,(2s)^2\,(2p)^q$ and $(1s)^2\,(2p)^{q+2}$, taking the $(2p)$ radial wave function the same for both configurations, have been worked out in detail, and numerical results evaluated for O^{+2} (Hartree, Hartree and Swirles 1939). For the lower state the ratio of the coefficients of the normalized wave functions, for the "excited" and "normal" configurations was about 0·2, indicating considerable departure from a single configuration; the effect on the radial wave functions was, however, found to be small, and not worth the considerable extra labour needed to evaluate these functions with superposition of configurations. A further development, taking the same two configurations but different $P(2p)$ radial wave functions for the two configurations, has also been examined (Jeffreys 1942), with much the same conclusion.

(b) *Polarization*

Consider one electron outside a core of closed groups, as for example, in neutral Na. The self-consistent field takes into account the *average* effect of an electron in one wave function on the other electrons in an atom, but does not include any representation of the "image force" on an outer electron arising from polarization of the core by it. For any position of the outer electron, well outside the closed groups, the induced dipole moment of the core will be proportional to $1/r^2$, and the attractive force on the outer electron will be proportional to $1/r^5$; its effect would be expressed by an additional contribution to V, proportional to $1/r^4$ for large distances (see, for example, Van Vleck and Whitelaw 1933).

Calculated values of some atomic properties, particularly transition probabilities, are rather sensitive to the wave functions from which they are calculated, so that it is necessary to use the best available wave functions from which to calculate them. In such cases it would be advisable to include polarization in calculating the wave functions for the outer electron.

A study of the polarization effect from the point of view of second-order perturbation theory has been made by Bethe (1933), and the subject has been

considered by Coolidge and James (1936) for the $(1s)(2s)^1S$ state of He, by James and Coolidge (1936) for Li, and by Eriksson (1940) for the P states of He. But a more generally applicable treatment is required; it would be an advantage if it were simpler and less formal, though such a treatment would probably also be less accurate.

Various attempts have been made to include, in a more or less empirical way, a "polarization" term in the potential of the field on an outer electron; these have mainly been in connection with configurations of one wave function outside complete groups. The argument at the beginning of this section suggests that this "polarization" potential should behave like $\frac{1}{2}\alpha/r^4$ for large r, but there is little indication how it should be taken to vary for small r.

In calculating the elastic scattering of electrons by argon and by krypton, Holtsmark (1929, 1930) used a polarization potential $\frac{1}{2}\alpha_e/r^4$ for large r, where α_e is the polarizability deduced from the observed dielectric constant, modifying this in an unspecified way for small r. Bierman and Harting (1942) and Bierman (1943) have used a polarization potential of the from $\frac{1}{2}(\alpha/r^4)[1-\exp\{-(r/r_0)^p\}]$ with $p=5$ and with $p=8$, and α and r_0 adjustable to give the best fit with optical term values; but since wave functions were calculated without exchange, the contribution to the energy ascribed to polarizability really includes a contribution from "exchange". Further, this formula gives a repulsive force for small r, which seems somewhat unlikely.

It is possible that a formula for the polarization potential could be found so that the ionization energies of the various terms agree closely with the observed terms of the optical spectrum, and if this could be done, it seems likely that the wave functions so obtained would be the best for the calculation of transition probabilities. But if an attempt is to be made to obtain such a formula, it would seem important to take into account as well as possible the other forms of interaction between the core and the outer electrons, and in particular to include exchange terms in the equations for the outer wave function.

Bates (1947) has used a polarization potential $\frac{1}{2}\alpha/(a^2+r^2)^2$ in an examination of the influence of polarization on the distribution of intensity in the continuous absorption spectrum of potassium beyond the series limit, and finds the results very sensitive to the "polarizability" α.

(c) *Relativistic wave functions*

In most of the work so far carried out, the one-electron wave functions Ψ taken have been of the form (2) and the Hamiltonian H used in applying the variation principle has been that for point electrons; electron spin has been included in a purely formal way by the spin wave function χ in (2). A few calculations have been made of doublet separations, taking the spin-orbit interaction as a perturbation (Hartree, Hartree and Manning 1941); but it has not been included in the calculation of the wave functions. Fock (1933) has considered the relativistic treatment of one wave function outside a configuration of closed groups, including exchange, and has given a justification of the formula for the spin doublet separation usually derived by means of a "spin-orbit interaction" term in Schrödinger's equation (cf. Condon and Shortley 1935, Ch. 5, § 4).

Although the correct relativistic form of quantum mechanics for a many-particle system is unknown, the form of the relativistic wave functions for one

electron in a central field is known, and it is possible to build up out of such relativistic one-electron wave functions an approximate wave function for the whole atom; some properties of such sets of relativistic wave functions have been considered by Hartree (1930). Further, it is possible to formulate a variation principle appropriate to such relativistic wave functions, and so to derive equations for the "best" one-electron wave functions to use in such an approximation to the wave function for the whole atom. Each one-electron wave function has four components, involving two radial wave functions and spherical harmonics; the integral for the total energy E can be expressed in terms of integrals over the radial wave functions only, and equations for the "best" radial wave functions can then be obtained by application of the variation principle (Swirles 1935—see the appendix of this Report for the author's corrections to this paper—and 1936). The only quantitative calculations are those of Williams (1940), for the normal state of Cu^+ without exchange (the necessary corrections appear to have been made in this work).

§ 7. ANALYTIC WAVE FUNCTIONS

Instead of leaving the form of the radial wave functions free so that the "best" functions can be determined from the variation principle, it is possible to restrict the radial wave functions contemplated to particular analytical forms with certain adjustable parameters, whose "best" values are then determined by appeal to the variation principle; that is to say, we attempt to find the "best" wave functions within the sets specified by the particular analytical forms adopted.

The main types of radial wave functions considered in this treatment are:

(i) products of exponentials and polynomials (Zener 1930);

(ii) sums of terms of the form $a_n r^n \exp(-\alpha_n r)$.

Tables for use with functions of the form (ii) and some examples of their use have been given by Morse, Young and Haurwitz (1935); the tables have been extended by Goldberg and Clogston (1939) and improved results given by Duncanson and Coulson (1944). They have mainly been applied to the light atoms; for heavier atoms the number of parameters becomes rather large, and the process of minimizing E with respect to the whole set of parameters becomes rather formidable. To keep down the number of parameters it is usual to use hydrogen-like wave function for $(2p)$; this involves only one parameter. This is only a rough approximation for atoms, not highly ionized, for which the $(2p)$ group is the outer group; if the parameter is chosen to give the right radius for the maximum of $P(2p; r)$, this function decreases too rapidly for large r (see figure 1). The restriction to this rather rough approximation may be a rather serious drawback to the use of analytical wave functions in contexts in which the behaviour of the wave function for large r is important. The same applies to the use of a hydrogen-like $(3d)$ wave function, as adopted by Goldberg and Clogston.

Wu and Shen (1945) have obtained analytical wave functions for some doubly-excited states of two- and three-electron systems.

For some applications, analytical wave functions are more convenient than functions specified by tables; for example, for evaluating the two-centre integrals

that arise in connection with applications to interaction between two atoms, and for the Fourier transform to wave functions in momentum space which are required for the evaluation of the shape of the Compton line in x-ray scattering (Dumond and Kirkpatrick 1937, Hicks 1937), and of inelastic scattering (Hughes and Mann 1938). Slater (1932) has shown that good approximations to the radial wave functions obtained by the solution of the equations of the self-consistent field can be found by use of linear combinations of terms of the type $r^m \exp(-a_m r)$. Another type of analytical wave function, considered as hydrogen-like wave functions modified by a scale change which is a function of r, has been examined by Vinti and Morse (1933).

Gombás (1941, 1942) has suggested an analytical treatment of the "self-consistent field" approach by approximating to the atomic field by an analytical formula and then using analytical wave functions for an electron in this field. This treatment has been applied by Kozma and Kónya (1941).

§ 8. WAVE FUNCTIONS INVOLVING r_{ij}

The characteristic feature of the approximate wave functions so far considered is that they are formed of products of the coordinates of the individual electrons only. For the simple case of the normal state of neutral He, the wave function of this kind is simply $f(r_1)f(r_2)$. Both on physical and analytical grounds it is clear that the wave function cannot be of this form, but must depend also on the mutual distance r_{12} between the electrons; an alternative third coordinate is θ_{12}, the angle between the vectors $\mathbf{r}_1$ and $\mathbf{r}_2$, or $\cos\theta_{12}$ given by $\cos\theta_{12}=(r_{12}^2-r_1^2-r_2^2)/2r_1r_2$. Similarly for a many-electron atom, the wave function must be a function of the mutual distances r_{ij} between all pairs of electrons. Improved approximation to the wave function would be obtained by including some dependence on the r_{ij}'s, but this means giving up the concept of the structure as built up out of a number of electrons each in a stationary state in the field of the nucleus and the space-charge distribution of the other electrons.

Most of the work on improved approximations of this kind has been done using analytical wave functions, by Hylleraas (1928, 1930, 1932) and others (Hylleraas and Undheim 1930, Hartree and Ingham 1933) for helium-like atoms and by James and Coolidge (1936) for lithium. In his earliest work Hylleraas (1928) used the coordinates r_1, r_2, θ_{12}; but in all the later work referred to, the coordinates used are r_1, r_2, r_{12}. Hylleraas has shown that it is practicable in this way to obtain a calculated value for the ionization energy of neutral helium in agreement with the experimental value to an accuracy of about 1 in 10^4; to this accuracy, relativity effects are appreciable, even for so light an atom as helium, and the finite mass of the nucleus also has to be taken into account.

In this work, there is one curious feature which seems to merit further examination. For the normal state of helium, let

$$r_1+r_2=ks,\ r_1-r_2=kt,\ r_{12}=ku.$$

The kind of wave function used by Hylleraas for the normal state is

$$e^{-ks}\Sigma_{nlm}a_{nlm}s^n t^l u^m;$$

the best wave function of this kind with a given finite set of terms of the sum is given by evaluating E in terms of k and the coefficients a_{nlm} and minimizing E with respect to these quantities. It was shown, however, by Bartlett, Gibbons

and Dunn (1935) that the actual wace function for the normal state of helium can certainly not be of the form

$$\Psi = \Sigma_{nlm} b_{nlm} s^n t^q u^m,$$

since if relations between the coefficients b_{nlm} in such an expansion are found by substituting it into the differential equation and equating coefficients, they are found to be mutually inconsistent. James and Coolidge (1937) showed how nevertheless the approximation to E could be improved indefinitely by taking further terms of such a series. They also showed how, if Φ is the actual wave function and Ψ an approximation to it, $\int[\Phi - \Psi]^2 d\tau$ could be estimated although Φ is not known, and that the value of this quantity, which is one measure of the goodness of an approximate wave function, decreases only slowly as the number of the terms of the series is increased, despite the considerable improvement in the energy value. This illustrates the limitations of the energy value as a criterion of an approximate wave function, and they discuss other possible criteria.

In connection with the wave function of the normal state of neutral helium, reference should be made to the work of Gronwall (1937), left unfinished at his death and later edited by Bartlett, and to Bartlett's (1937) extension of it.

Work on the excited states of He, taking the r_{12} term into account, has been done by Hylleraas and Undheim (1930) and more recently by Eriksson (1942) Coolidge and James (1936) have made a critical examination of various approximate wave functions for the $(1s)(2s)\,^1S$ state of He.

Baber and Hassé (1937) attempted to find an approximate wave function for the normal state of neutral helium of the form

$$\Psi = \psi(r_1)\psi(r_2)\phi(r_{12})$$

without restricting the analytical form of the functions ψ and ϕ, but on account of numerical difficulties were driven to adopt a specific analytical form for ϕ, Stevenson (1937 b) has suggested a rather similar treatment, using r_1, r_2 and θ_{12} as coordinates of the outer wave functions of a configuration of $(ns)^2$ outside closed groups.

§ 9. RESULTS

(a) *Low states of ionization and excitation*

Most of the results which have been calculated are for normal states or states of low excitation (giving terms of the optical spectrum) of atoms in states of ionization from -1 to $+4$. A summary of these results is given in table 1 in a tabular form somewhat similar to that adopted by Shenstone (1938) in an earlier volume of these Reports (Vol. V, p. 222) in cataloguing the analysed optical spectra of the various states of ionization of the atoms.

Each row refers to a single element and each column to a single degree of ionization, so that iso-electronic systems lie on diagonal lines.

The entries show briefly the kind of information available, by means of the following code:—

S indicates wave functions calculated by the method of the self-consistent field;

A indicates analytical wave functions with parameters determined by the use of the variation principle;

X indicates that exchange terms have been included in the determination of the wave functions;

R indicates relativistic wave functions;
N indicates the normal configuration;
E indicates singly-excited configurations;
D indicates doubly-excited configurations;
C indicates states of the continuum;
M indicates wave functions for the whole atomic system involving the mutual distances r_{ij} (analytical wave functions except in the case of Baber and Hassé (1937));
H indicates one-electron systems for which exact hydrogen-like wave functions are available;
F indicates that fields only, not wave functions, are available.

The letter and number below give the reference in the bibliography. For example NEAX M4 indicates that analytical wave functions for the normal and some excited states have been calculated including exchange terms, and the results are to be found in reference M4 of the bibliography. For the results indicated by E or D, reference must be made to the papers quoted to find for just what excited states wave functions have been calculated.

The effect of the inclusion of exchange terms on the radial wave functions of the inner groups of atoms is not large, but the effect on the outer groups may be considerable, particularly if their $\epsilon_{nl,\,nl}$ values are small (roughly speaking, they are "loosely bound") as for the $(3d)^{10}$ group of Cu, the $(3p)^6$ group of Cl^-, or generally the outermost group of a negative ion. In most cases of one electron outside closed groups, giving the states corresponding to the terms of one-electron spectra such as Na^+ and Ca^{+2}, the exchange terms form a comparatively small contribution to $d^2P(nl;\,r)/dr^2$ over most of the range, and their effect on $P(nl;\,r)$ is not large. But in a few cases, of which $Ca^+(3d)$ is one example, the exchange terms form the main contribution to $d^2P(nl;\,r)/dr^2$ in the neighbourhood of the main maximum, so that the results calculated with and without exchange are quite different; for $Ca^+(3d)$ the radius of the main maximum of P is altered by a factor 2 by the exchange terms (see D. R. and W. Hartree 1938, § 5), and transition probabilities depending on this wave function are quite different according as it was calculated with or without exchange.

Inclusion of exchange terms also leads to different radial wave functions for the different terms arising from a single configuration. In Slater's (1929) theory of complex spectra, the argument leading to the formulae for the contributions to the energy from incomplete groups is based on the application of perturbation theory to the set of wave functions $\Psi(A;\,R)$ corresponding to different assignments of $(m_l,\,m_s)$ values to the one-electron wave functions not in complete groups, these functions $\Psi(A;\,R)$ being treated as degenerate. The different terms are then represented by different linear combinations of the same set of determinations $\Psi(A;\,R)$, and this implies the use of the *same* radial wave functions for all terms arising from a single configuration.

But once the linear combinations corresponding to the different terms, and the corresponding energy formulae in terms of the radial wave functions have been established, by this or any other argument, this restriction can be removed, and indeed it has to be removed if the energy for each term is to be minimized eparately.

Table 1. Summary of results of calculations of atomic structures

State of ionization / Spectrum		−1	0 I	+1 II	2 III	3 IV	4 V
Atom	Atomic number N						
H	1	NM C H24 M2 W3 C1–4	H				
He	2	C M1 M2	NM EM NEDS NEAX H18 H28 W4 M7 H27 H29 W5 G2 H28 H30 W1 H29 E1	H			
Li	3		NM NSX ES NEAX NAX DAX J1 F3 H2 M7 D2 W7 G2 W8	NM NS NEAX H29 H2 M7 G2	H		
Be	4		NS NSX ESX NEAX NAX H8 H9 H10 M7 D2 F2	NEAX M7 G2	NS NEAX H8 M7 G2	H	
B	5		NS NAX B4 M7 D2	NEAX M7	NEAX M7 G2	NEAX M7 G2	H
C	6		NS NSX NES NAX B4 J2 T1 D2 U1	NAX M7	NSX NEAX J2 M7	NEAX M7 G2	NSX NEAX J2 M7 G2
N	7	NSX H16	NS NSX NAX B4 H16 D2		NAX M7	NEAX M7	NEAX M7 G2
O	8	NSX H23	NS NSX NAX H17 H23 D2	NS NSX H17 H23	NS NSX H17 H23	NS NSX NAX H17 H23 M7	NEAX M7

F	9	NS NAX B3 M7 H7	NS NAX B3 D2				NAX M7
Ne	10		NS NAX B3 M7 D2				
Na	11	NSX H15	NSX F1 H16	NSX NAX F1 M7 H16			
Mg	12				NS NSX NAX B1 Y1 M7 Y2		
Al	13			NS B1		NS H7	
Si	14				NS D1	NS NES NESX D1 M3 H22	NS NSX M3 H22
P	15						
S	16						
Cl	17	NS NSX C H4 H11 H19	NS H19				
Ar	18	C H25	NFS NSX H25 H14				
K	19	NSX H15	NS G1	NS NSX H5 H14			
Ca	20		NS NESX H8 H13	NESX H13	NS NSX H8 H13		
Sc	21				NS H1	NS H1	

Table 1 (*cont.*)

State of ionization / Spectrum		−1	0 I	+1 II	2 III	3 IV	4 V
Atom	Atomic number N						
Ti	22					NS H1	NS H1
V	23						
Cr	24		NS P1		NS P1 M6		
Mn	25						
Fe	26		NS M4				
CO	27						
Ni	28		NS G3				
Cu	29			NS NSX NRS H4 H12 W2			
Zn	30		NS H20				
Ga	31		NS H20	NS H20		NS H20	

Ge	32		NS H21		NS H21		
As	33		NS H20	NS H20	NS H20	NS H20	
Se	34						
Br	35	C A2	NFS A2				
Kr	36	C H26 A1	NFS H26				
Rb	37			NS H7			
Sr	38						
Yt	39						
Zr	40						
Ag	47			NS B2			
Cs	55			NS H5			
W	74		NS M5				
Hg	80		NS H8		NS H6 H8		

The difference between the radial wave functions for different terms arising from a single configuration is greatest for the outermost (nl) group, and may be quite considerable for a neutral atom (see D. R. and W. Hartree 1936 a, figure 1 and 1938 a, figure 2).

(b) *Negative ions*

Attempts to calculate the structures of negative ions, say by the method of the self-consistent field with or without exchange, may raise some points which do not arise in similar work for neutral atoms or positive ions, in that the negative ion considered may not be able to exist. There are two different ways in which the results may indicate that, to the approximation being used, the negative ion could not exist as a permanent structure. On the one hand, it may be found that the equations had no solution at all, so that in that approximation the negative ion could not exist at all. On the other hand, it may be possible to obtain a solution for a singly negative ion, but the total energy of the negative ion may be greater than that of the neutral atom. In such a case the negative ion is a possible structure but would be liable to auto-ionization, that is, spontaneous dissociation into a neutral atom and a free electron with positive energy; its existence would have a transitory character of the same nature as, for example, that of the doubly-excited $(2s)^2$ state of He, which has a greater energy than that of $He^+(1s)$ and has a high probability of dissociation into He^+ and a free electron. For a negative ion this may happen even in the normal state, and may be expressed by speaking of the ion as "unstable". A similar argument applies to higher states of negative ionization.

If, on the basis of any particular approximation, a solution of the equations for a negative ion can be found, it does not seem possible to draw quite definite conclusions about the stability of the actual ion from calculation of energies. If E^0 is the actual value of the *total* energy of the normal state of a neutral atom and E^- is that of the negative ion, and E_c^0, E_c^- are the calculated values of these energies on the basis of some approximate form of wave function, the ionization energy of the negative ion (electron affinity of the neutral atom) is $E^0 - E^-$; if this is positive, the ion is stable. The corresponding calculated ionization energy is $E_c^0 - E_c^-$. Now the variation principle only states that the *total* energy of a normal state, calculated from any approximate wave function, is greater than the actual *total* energy, that is, in this case, $E_c^0 > E^0$ and $E_c^- > E^-$, and not that a similar inequality holds for each *ionization* energy. Thus of the four arrangements of the relative positions of E, E_c, E_c^-, and E_c^-, illustrated in figure 2, no one is excluded by the variation principle; and if all four cases could occur, nothing about the stability or instability of the actual negative ion could be deduced from the calculated energies. However, case IV of figure 2 can only occur if $(E_0^- - E^-) < (E - E^0)$, that is, of the error in the *total* energy *decreases* with increasing number of electrons. It seems likely that if a consistent type of approximate wave function is used throughout, the error in the calculated value of the total energy will increase with the number of electrons, so that case IV of figure 2 will not arise. Then if the calculated energies indicate a stable negative ion, it can be concluded that the actual negative ion is stable. But this conclusion must be regarded as probable and not certain. For the same reason, case II is not unlikely to occur, so that no conclusion can be drawn from calculated energies

which would indicate an unstable negative ion, unless some estimate of the error in the calculated ionization energy can be made.

Results for negative ions have been calculated by the method of the self-consistent field with exchange for N^-, O^-, Na^- and K^-. The calculated ionization energy is positive for O^-, but negative (indicating an unstable ion, to the accuracy

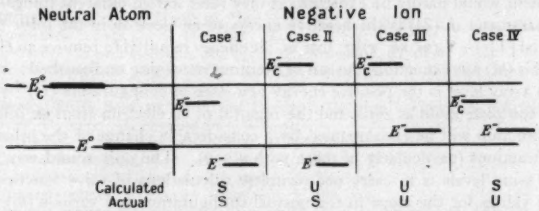


Figure 2. Relative position of energy levels of normal states of neutral atom and negative ion.
U indicates unstable negative ion.
S ,, stable ,, ,,

of the approximation adopted) for the other three ions. However, consideration of the probable errors of the calculated energies, due to the approximation used, suggest that Na^- and K^- are probably both stable (D. R. and W. Hartree 1938). The stability of N^- seems quite uncertain.

A rather thorough attempt has been made to obtain a solution of Fock's equations for O^-[2], but without success. To the accuracy of the approximation on which these equations are based, it seems unlikely that this ion can exist in the free state (Hartree, Hartree and Swirles 1939, p. 236). A similar negative result has been found by Massey in an attempt to obtain a solution for an excited state $(2p)^4(3s)$ of O^-.

The H^- ion has been studied by Williamson (1942) and later by Chandrasekhar (1945), using wave functions of Hylleraas' type depending on r_{12} as well as on r_1 and r_2.

(c) *High states of ionization*

Results have been calculated for a few higher states of ionization, mainly for two and three-electron structures (NEAX in the code introduced above, references M4, G2, C2), Cr^{+6} (NS, reference P1), Fe^{+16} to Fe^{+13} (NSX, Gold, not yet published). The high states of ionization of Fe are of interest in connection with Edlén's (1942) identification of some of the strong lines in the solar corona as being due to these ions.

(d) *States of high excitation*

A few results have been calculated for states of high excitation giving terms of the x-ray spectra, for Na by Kennard and Ramberg (1934) and for K by Thatcher (1939). The wave functions were calculated without exchange, but exchange terms were included in the calculation of the energy; this procedure is valid in view of the stationary property of the energy, calculated from (2), for small variations in the wave function. Some doubly-excited states were calculated by Kennard and Ramberg, in connection with the interpretation of satellite lines in x-ray spectra.

Already in the first work on self-consistent field calculations (Hartree 1928), the close agreement between the $\epsilon_{nl,\,nl}$ parameters of equations (24) and the energies of the x-ray terms was noted. But, as was recognized at the time and as emphasized by others (Gaunt 1928, Fock 1933, Koopmans 1933), although there is a correspondence between these two sets of quantities, close quantitative agreement would hardly be expected, as they refer to two different things. The $\epsilon_{nl,\,nl}$ parameter in (24) is the negative energy of an electron in the *static* field of potential $[\mathrm{T}(r)+\mathrm{Y}_0(nl, nl; r)]/r$, that is, the energy required to remove an electron from this (nl) wave function, the rest of the atom remaining undisturbed; whereas an (nl) x-ray level is the positive energy of a state of configuration $(nl)^{-1}$ relative to the complete atom as zero, and the removal of an electron from an inner (nl) wave function will be accompanied by a considerable change of the other $(n'l')$ wave functions (particularly of those with $n'>n$). The only sound way to calculate x-ray levels is to carry out complete calculations of wave functions and energy values for the atom in the normal configuration and various $(nl)^{-1}$ configurations and evaluate the appropriate energy differences. This has been done by Thatcher (1939) for the normal state of K^+ and the $(1s)^{-1}$ and $(2p)^{-1}$ states of K^{+2}. The agreement of the calculated x-ray levels with those observed is, however, rather worse than the agreement of the $\epsilon_{nl,\,nl}$ parameters with the observed levels.

(e) *States of the continuum*

Calculations of the self-consistent field type without or with exchange, can be extended to the continuum of positive-energy states. Those which have been carried out are indicated by C in table 1. The "degree of ionization" is reckoned as that of the whole system of (atom + electron in a state of positive energy); for example, the states of the eleven-electron system of a Na^+ ion and a free electron are classed as states of the continuum of neutral Na. The continuous absorption spectrum of Na is then ascribed to transitions between discrete states and states of the continuum of neutral Na; and for consistency, the scattering of electrons by Ne is expressed in terms of wave functions of Ne^-.

Calculations for the continuum differ in some respects from those for bound states. First, the energy of the free electron, which takes the place of the $\epsilon_{nl,\,nl}$ parameter which for bound states has to be determined so that the solution has the required behaviour, is normally a *given* parameter; this is a great simplification. Secondly, the fact that the radial wave function P remains finite as $r\to\infty$, instead of decreasing exponentially to zero, calls for different treatment of the equation for large r. One possible treatment has been given by Hartree, Kronig and Petersen (1934). Thirdly, for some applications the only results wanted are the "phases" η_l of the asymptotic solution for large r, defined by

$$\mathrm{P}(El;\ r)\sim \mathrm{A}\sin\left(\sqrt{(2E)}r-\tfrac{1}{2}l\pi+\eta_l\right).$$

Only these phases are required for calculating elastic scattering; but for calculating transition probabilities to the continuum, either by absorption of radiation (see for example Bates 1947) or by electron collision, the wave function for small r is also required. The wave functions for the continuum of H^{-1} have been studied by Williamson (1942), Henrich (1944) and Chandrasekhar (1944, 1945, 1946) in connection with the continuous absorption spectrum of H^- and its significance in connection with the solar spectrum.

REFERENCES

A 1 ARNOT, F. L. and BAINES, G. O., 1934. *Proc. Roy. Soc.*, A, **146**, 651.
A 2 ARNOT, F. L. and MCLAUCHLAN, T. C., 1934. *Proc. Roy. Soc.*, A, **146**, 662.
BABER, T. D. and HASSÉ, H. R., 1937. *Proc. Camb. Phil. Soc.*, **33**, 253.
BARTLETT, J. H., 1937. *Phys. Rev.*, **51**, 661.
BARTLETT, J. H., GIBBONS, J. J. and DUNN, C. G., 1935. *Phys. Rev.*, **47**, 79.
BATES, D. R., 1947. *Proc. Roy. Soc.*, A, **188**, 350.
BETHE, H. A., 1933. *Handb. d. Phys.*, **24**/1, Ch. 3, p. 339.
BIERMANN, L., 1943. *Z. Astrophys.*, **22**, 157.
B 1 BIERMANN, L. and HARTING, H., 1942. *Z. Astrophys.*, **22**, 81.
B 2 BLACK, MURIEL M., 1935. *Mem. and Proc. Manchester Lit. and Phil. Soc.*, **79**, 29.
B 3 BROWN, F. R., 1933. *Phys. Rev.*, **44**, 214.
B 4 BROWN, F. R., BARTLETT, J. H. and DUNN, C. G., 1933. *Phys. Rev.*, **44**, 296.
BUSH, V., 1931. *J. Franklin Inst.*, **212**, 447.
CALDWELL, S. H., 1932. *Phys. Rev.*, **42**, 589.
C 1 CHANDRASEKHAR, S., 1944. *Astrophys. J.*, **100**, 176.
C 2 CHANDRASEKHAR, S., 1945 a. *Astrophys. J.*, **102**, 223.
C 3 CHANDRASEKHAR, S., 1945 a. *Astrophys. J.*, **102**, 395.
C 4 CHANDRASEKHAR, S. and BREEN, FRANCES H., 1946. *Astrophys. J.*, **103**, 41.
CONDON, E. V. and SHORTLEY, G. H., 1935. *Theory of Atomic Spectra* (Cambridge : University Press).
C 5 COOLIDGE, A. S. and JAMES, H. M., 1936. *Phys. Rev.*, **49**, 676.
DELBRÜCK, 1930. *Proc. Roy. Soc.*, A, **129**, 686.
DIRAC, P. A. M., 1940. *Proc. Camb. Phil. Soc.*, **26**, 276 ; 1931. *Ibid.*, **27**, 240.
D 1 DONLEY, H. L., 1936. *Phys. Rev.*, **50**, 1012.
DUMOND, J. and KIRKPATRICK, H. A., 1937. *Phys. Rev.*, **52**, 419.
D 2 DUNCANSON, W. E. and COULSON, C. A., 1944. *Proc. Roy. Soc. Edinb.*, **62**, 37.
ECKERT, C., 1930. *Phys. Rev.*, **36**, 878.
EDLEN, B., 1942. *Z. Astrophys.*, **22**, 30.
ERIKSSON, A., 1942. *Nova Acta Soc. Sci. upsal.*, **11**, No. 9.
FOCK, V., 1930. *Z. Phys.*, **61**, 126 ; *Ibid.*, **62**, 795 ; 1933. *Ibid.*, **81**, 195.
F 1 FOCK, V. and PETRASHEN, MARY J., 1934. *Phys. Z. Sowjet*, **6**, 368.
F 2 FOCK, V. and PETRASHEN, MARY J., 1935 a. *Phys. Z. Sowjet*, **8**, 359.
F 3 FOCK, V. and PETRASHEN, MARY J., 1935 b. *Phys. Z. Sowjet*, **8**, 457.
FRANK, N. H., 1937. *Phys. Rev.*, **51**, 577.
GAUNT, J. A., 1928. *Proc. Camb. Phil. Soc.*, **24**, 328.
G 1 GIBBONS, J. J. and BARTLETT, J. H., 1935. *Phys. Rev.*, **47**, 692.
G 2 GOLDBERG, L. and CLOGSTON, A. M., 1939. *Phys. Rev.*, **56**, 696.
GOMBÁS, P., 1941. *Z. Phys.*, **118**, 164 ; 1942. *Ibid.*, **119**, 318.
G 3 GRAY and MANNING, M. F., 1941. *Phys. Rev.*, **59**, 475.
GRONWALL, T. H., 1937. *Phys. Rev.*, **51**, 655.
H 1 HANCOCK, S. T. R., 1934. *M.Sc. Thesis*, Manchester.
H 2 HARGREAVES, J., 1928. *Proc. Camb. Phil. Soc.*, **25**, 75.
H 3 HARTREE, D. R., 1927. *Proc. Camb. Phil. Soc.*, **24**, 89, 111.
HARTREE, D. R., 1929. *Phys. Z.*, **30**, 517 ; 1930. *Proc. Camb. Phil. Soc.*, **25**, 225 ; 1933 a. *Mem. and Proc. Manchester Lit. and Phil. Soc.*, **77**, 91.
H 4 HARTREE, D. R., 1933 b. *Proc. Roy. Soc.*, A, **141**, 282.
H 5 HARTREE, D. R., 1934 a. *Proc. Roy. Soc.*, A, **143**, 506.
H 6 HARTREE, D. R., 1934 b. *Phys. Rev.*, **46**, 738.
H 7 HARTREE, D. R., 1935. *Proc. Roy. Soc.*, A, **151**, 96.
HARTREE, D. R., 1938. *Proc. Camb. Phil. Soc.*, **25**, 225.
H 8 HARTREE, D. R. and W., 1935 a. *Proc. Roy. Soc.*, A, **149**, 210.
H 9 HARTREE, D. R. and W., 1935 b. *Proc. Roy. Soc.*, A, **150**, 9.
H 10 HARTREE, D. R. and W., 1936 a. *Proc. Roy. Soc.*, A, **154**, 588.
H 11 HARTREE, D. R. and W., 1936 b. *Proc. Roy. Soc.*, A, **156**, 45.
H 12 HARTREE, D. R. and W., 1936 c. *Proc. Roy. Soc.*, A, **157**, 490.
H 13 HARTREE, D. R. and W., 1937. *Proc. Roy. Soc.*, A, **164**, 167.
H 14 HARTREE, D. R. and W., 1938 a. *Proc. Roy. Soc.*, A, **166**, 450.
H 15 HARTREE, D. R. and W., 1938 b. *Proc. Camb. Phil. Soc.*, **34**, 550.

H16 Hartree, D. R. and W., 1948. *Proc. Roy. Soc.*, A (in press).
H17 Hartree, D. R. and Black, M. M., 1933. *Proc. Roy. Soc.*, A, **139**, 311.
H18 Hartree, D. R. and Ingham, A. C., 1933. *Mem. and Proc. Manchester Lit. and Phil. Soc.*, **77**, 69.
H19 Hartree, D. R., Kronig, R. de L. and Petersen, H., 1934. *Physica*, **1**, 895.
Hartree, D. R. and Swirles, B., 1937. *Proc. Camb. Phil. Soc.*, **33**, 240.
H20 Hartree, D. R. and W. and Manning, M. F., 1941 a. *Phys. Rev.*, **59**, 299.
H21 Hartree, D. R. and W. and Manning, M. F., 1941 b. *Phys. Rev.*, **59**, 306.
H22 Hartree, D. R. and W. and Manning, M. F., 1941 c. *Phys. Rev.*, **60**, 857.
H23 Hartree, D. R. and W. and Swirles, B., 1939. *Phil. Trans. Roy. Soc.*, A, **238**, 229.
Hasse, H. R., 1930. *Proc. Camb. Phil. Soc.*, **26**, 542.
H24 Henrich, L. R., 1944. *Astrophys. J.*, **99**, 59.
Hicks, B., 1937. *Phys. Rev.*, **52**, 436.
H25 Holtsmark, 1929. *Z. Phys.*, **55**, 437.
H26 Holtsmark, 1930. *Z. Phys.*, **66**, 49.
Hughes, A. L. and Mann, M. M., 1938. *Phys. Rev.*, **53**, 50.
H27 Hylleraas, E. A., 1928. *Z. Phys.*, **48**, 469.
H28 Hylleraas, E. A., 1929. *Z. Phys.*, **54**, 347.
Hylleraas, E. A., 1930. *Z. Phys.*, **65**, 209 ; *Ibid.*, **66**, 453.
H29 Hylleraas, E. A., 1932. *Norske Videnskaps-Akademi Skifter*; I, *Math. Naturwiss. Klasse*, No. 6.
H30 Hylleraas, E. A. and Undheim, B., 1930. *Z. Phys.*, **65**, 759.
James, H. M., 1936. *Phys. Rev.*, **49**, 874 ; 1941. *Bull. Amer. Math. Soc.*, **47**, 869.
James, H. M. and Coolidge, A. S., 1935. *Phys. Rev.*, **47**, 700.
J1 James, H. M. and Coolidge, A. S., 1936. *Phys. Rev.*, **49**, 688.
James, H. M. and Coolidge, A. S., 1937. *Phys. Rev.*, **51**, 855, 860.
Jeffreys, Bertha, 1942. *Proc. Camb. Phil. Soc.*, **38**, 290.
J2 Jucys, A., 1939. *Proc. Roy. Soc.*, A, **173**, 59.
Kennard, E. H. and Ramberg, E., 1934. *Phys. Rev.*, **46**, 1034, 1040.
Kimball, G. E. and Shortley, G. H., 1934. *Phys. Rev.*, **45**, 815.
Koopmans, T., 1933. *Physica*, **1**, 104.
Kozma, B. and Kónya, A., 1941. *Z. Phys.*, **118**, 153.
Kramers, H. A., 1938. *Handb. u. Jahrb. Chem Phys.* I, Sect. 7, p. 313.
Lennard-Jones, J. E., 1931. *Proc. Camb. Phil. Soc.*, **27**, 469.
M1 McDougall, J., 1932 a. *Proc. Camb. Phil. Soc.*, **28**, 341.
M2 McDougall, J., 1932 b. *Proc. Roy. Soc.*, A, **136**, 549.
M3 McDougall, J., 1932 c. *Proc. Roy. Soc.*, A, **138**, 550.
M4 Manning, M. F. and Goldberg, L., 1938. *Phys. Rev.*, **53**, 662.
M5 Manning, M. F. and Millman, J., 1936. *Phys. Rev.*, **49**, 848.
Millard and Manning, M. F., 1938.
M6 Mooney, R. L., 1934. *Phys. Rev.*, **55**, 557.
M7 Morse, P. M., Young, L. A. and Haurwitz, Eva S., 1935. *Phys. Rev.*, **48**, 948.
P1 Porter, A., 1934. *Mem. and Proc. Manchester Lit. and Phil. Soc.*, **79**, 75.
Shenstone, A. G., 1938. *Reports on the Progress of Physics*, Vol.V (Physical Society, London).
Shortley, G. H., 1936. *Phys. Rev.*, **50**, 1072.
Shoupp, W. E., Bartlett, J. H. and Dunn, C. G., 1935. *Phys. Rev.*, **47**, 705.
Slater, J. C., 1938. *Phys. Rev.*, **32**, 339 ; 1929. *Ibid.*, **34**, 1293 ; 1930. *Ibid.*, **35**, 210 ; *Ibid.*, **36**, 57 ; 1932. *Ibid.*, **42**, 33.
Southwell, R. V., 1943. *Proc. Roy. Soc.*, A, **184**, 253 ; 1946. *Relaxation Methods in Theoretical Physics* (Oxford : University Press).
Stevenson, A. F., 1937 a. *Phys. Rev.*, **51**, 285 ; 1937 b. *Proc. Roy. Soc.*, A, **160**, 588 ; 1939. *Phys. Rev.*, **56**, 586.
Swirles, Bertha, 1935. *Proc. Roy. Soc.*, A, **152**, 625 ; 1936. *Ibid.*, **157**, 680.
Thatcher, W. A., 1939. *Proc. Roy. Soc.*, A, **172**, 242.
T1 Torrance, C., 1934. *Phys. Rev.*, **46**, 388.
U1 Ufford, C. W., 1938. *Phys. Rev.*, **53**, 568.
Van Vleck, J. H. and Whitelaw, N. G., 1933. *Phys. Rev.*, **44**, 551.
Vinti, J. P. and Morse, P. M., 1933. *Phys. Rev.*, **43**, 337.

VAN DER WAERDEN, 1932.
WALLER, I. and HARTREE, D. R., 1929. *Proc. Roy. Soc.*, A, **124,** 119.
W1 WHEELER, T. S., 1942. *Proc. R. Irish Acad.*, **48,** 43.
WHITE, H. E., 1934. *Introduction to Atomic Spectra* (McGraw-Hill).
W2 WILLIAMS, A. O., 1940. *Phys. Rev.*, **58,** 733.
W3 WILLIAMSON, R. E., 1942. *Phys. Rev.*, **62,** 538.
WILLS, L. A. and BREIT, G., 1935. *Phys. Rev.*, **47,** 704.
W5 WILSON, W. S., 1935. *Phys. Rev.*, **48,** 536.
W4 WILSON, W. S. and LINDSAY, R. B., 1935. *Phys. Rev.*, **47,** 681.
W7 WU, T-Y., 1940. *Phys. Rev.*, **58,** 1114.
W6 WU, T-Y. and MA, S. T., 1935. *Phys. Rev.*, **49,** 917.
W8 WU, T-Y. and SHEN, S. T., 1944. *Chinese J. Phys.*, **5,** 150.
YOST, W. J., 1940 a. *Phys. Rev.*, **57,** 557.
Y1 YOST, W. J., 1940 b. *Phys. Rev.*, **58,** 557.
ZENER, C., 1940. *Phys. Rev.*, **36,** 51.

APPENDIX

Corrections and additions to paper on "The Relativistic Self-consistent Field", by B. SWIRLES, *Proc. Roy. Soc.*, A, **152,** 625 (1935).

(i) The transformation properties of the functions ψ for different choices of the matrices α and β were not considered correctly, and § 2 is not internally consistent as it stands. The simplest correction is the following. Write $-\alpha_1$, $-\alpha_2$, $-\alpha_3$, $-\beta$ on the left of equations (2) and alter (18) accordingly so that it agrees with (7.1) of Darwin's paper quoted. Then equations (19 A), (19 B), (20 A), (20 B) stand, but (21) and (23) are modified. (23) becomes

$$\left(\frac{\epsilon+v}{c}+c\right)Q_l+\frac{dP_l}{dr_l}-\frac{l+1}{r}P_l=0$$

$$\left(-\frac{\epsilon+v}{c}+c\right)P_l+\frac{dQ_l}{dr}+\frac{l+1}{r}Q_l=0.$$

It is then true that P_l is the "large" and Q_l the "small" radial wave function; this is clearly not the case for the equations given in the paper. In § 3 the expression (32) requires corresponding changes of sign.

P. 629, delete footnote from "Differences of sign . . ." to the end.

(ii) An independent calculation of the coefficients had been made by Inglis (see Condon and Shortley, 1935, p. 261). The tables agree except in one place; on p. 641 of the paper, the second and third figures in the ninth column should be interchanged.

(iii) § 4. Table III and comment on p. 647. When relativistic terms are neglected, the *total* interactions within a complete group agree with those given by D. R. and W. Hartree, 1936 b, p. 48. See also Condon and Shortley, 1935, Chap. 6, § 9, in particular equations (2) and (11). The quantity evaluated is the same whether (n, l, m_l, m_s) or (n, l, j, m) wave functions are used. This being so, the terms in (57) and (58) arising from electrostatic interaction should be the same whether relativistic terms are neglected or included. That they are not is due to an error in differentiating. The corrections are :—

P. 648. The third line should read :—

$$\frac{dP_{2,1}}{dr}-\frac{2}{r}P_{2,1}+\left[c+\frac{\epsilon_{(2,1)}}{c}+\frac{Z}{cr}\right]Q_{2,1}-\frac{\epsilon_{(2,1\,;\,3,1)}}{c}Q_{3,1}.$$

In the eighth line read " 4/25 " for " 8/25 ".

The 13th line should read :—

$$\frac{dP_{2,-2}}{dr}+\frac{1}{r}P_{2,-2}+\left[c+\frac{\epsilon_{(2,-2)}}{c}+\frac{Z}{cr}\right]Q_{2,-2}-\frac{\epsilon_{(2,-2\,;\,3,-2)}}{c}Q_{3,-2}.$$

Omit the " boxes " round certain terms in the equations.

P. 649. Full stop after " same " in second line, and delete " except . . . equations ".

DEVELOPMENTS IN THE INFRA-RED REGION OF THE SPECTRUM

G. B. B. M. SUTHERLAND* AND E. LEE †,

* Pembroke College, Cambridge. † Royal Naval Scientific Service

§1. INTRODUCTION

THE developments which have taken place in the infra-red region of the spectrum during the past seven years will probably have a very great effect on the general development of science in the next decade. The speed and precision with which infra-red radiation can now be detected has made a knowledge of infra-red methods indispensable for scientists in a variety of fields. The organic chemist is gradually realizing that the most characteristic physical property of any chemical compound is its infra-red spectrum, which can now be obtained in little more time than it takes to determine a boiling point. The physical chemist has for some time realized the power of infra-red spectroscopy in determining atomic configurations and internuclear distances in simple molecules, but the method has now been extended to provide vital structural information (i.e. presence of key groups) about complex polyatomic molecules, such as penicillin and the giant polymers on which the plastics industry is based. As an analytical and structural tool in the investigation of matter, infra-red analysis is now taking its place alongside X-ray analysis for all except metallic substances. It has the advantage over the X-ray method that the material under investigation can be in the liquid or gaseous form and that even in the solid state, single crystals are not required, except for work with polarized radiation. In meteorology, the new techniques should allow long needed data to be obtained for the calculation of radiative equilibrium in the stratosphere, quite apart from the determination of the constitution of the higher layers of the atmosphere. In astrophysics, the new detectors may enable us to "see" farther and will certainly permit the study of the spectra of stars and planetary atmospheres in a much more comprehensive manner than was previously thought possible. The new ability to measure rapidly varying surface temperatures as low as 100° C. will be of great importance in the study of friction and of other problems of interest to engineers. In the biological field, the use of infra-red analysis is likely to be severely limited by the intense absorption due to water over most of the useful region. However, when the biologically active molecule can be isolated in a solid state, or in solution in a suitable non-aqueous solvent, the usual methods can be applied. In this connection, infra-red analysis may be a useful supplement to the tracer method of radioactive isotopes since nonradioactive isotopes can be detected by infra-red methods. Another potential application of the new technique to biological science lies in the field (hitherto virtually undeveloped) of infra-red microscopy. This last should be particularly valuable if polarized radiation is employed.

The advances in the physics of the infra-red which have made possible these applications in other fields are not purely technical, but have intrinsic interest

to physicists. The search for more sensitive and more rapid detectors has led to a re-consideration of the factors which govern the ultimate limits of detection in the infra-red. This has brought new results in the theory of temperature fluctuations. Again, the exact mechanism of the photoconductive (lead sulphide) cell is a problem of considerable importance in the theory of semi-conductors. Some of the wartime advances in technique were made by trial and error methods and it is hoped that a review of this kind may stimulate others to study the physics of the infra-red for its own sake, since it is certain that further advances will be made when we have a thorough understanding of the physics of infra-red detection.

We shall consider first the developments in the technique of detection as we proceed from the short-wave boundary of the infra-red to the long-wave boundary in the region of ultra-short radio waves. We follow this with an account of the new optical materials which have been developed for use in the infra-red as windows, filters and prisms. Finally, we deal with some of the applications, special attention being given to infra-red spectroscopy. Although a good deal of the war time work has now been published, many of the details are still not generally accessible so that a number of references are missing from this Report and sometimes only authors' names are given. The information on German work has been obtained from a number of sources some of which are not generally accessible, largely due to publication difficulties. We have tried to be fair in allocating credit for this work and hope that future publication by the authors themselves will correct any mistakes we may have made. It will also be noticed that the figures available for the performance of certain detectors are based on their wartime applications, rather than on absolute physical standards. It is felt, however, that it would be wrong to delay this report until all such information has been obtained in view of the extent and the importance of the available knowledge.

§ 2. DETECTORS

(i) *General considerations*

The recent advances in infra-red detection involve no new physical principle, but consist in developments of the traditional methods which may be conveniently listed as:—

(*a*) photochemical (photographic)
(*b*) phosphor
(*c*) photo-emissive
} useful to 1·3 μ.

(*d*) photoconductive
(*e*) photovoltaic
} useful to 5 μ.

(*f*) thermoelectric
(*g*) bolometric
(*h*) pneumatic
(*i*) radiometric
} useful over all wavelengths.

(*j*) evaporographic and convective—have been used to 10 μ.

The first five of these are often referred to as "selective" detectors, since they are sensitive only over a limited region in the infra-red, and even within that range have not a uniform response to all wavelengths. The last five are by

contrast called "non-selective" detectors since they respond equally to all wavelengths in the infra-red provided the receiving element is suitably blackened. This nomenclature is not entirely satisfactory because it does not indicate the real physical distinction between the two types of detector. The so-called "selective" detectors would in our opinion be more properly called *photo-detectors*, since in principle they react to the individual photons, whereas the "non-selective" detectors should be called *thermo-detectors*, since they react to the change in temperature of the detector caused by the absorption of radiation. The fact that temperature equilibrium must be achieved in the latter type of detector makes it fundamentally different from the former as we shall see when we come to consider the ultimate limits of detection.

The other general point we wish to stress before discussing individual detectors is the method of describing their performance. A frequent practice has been to give either the output (e.g. in microvolts) per microwatt of radiation, or the minimum detectable incident power in watts. In the case of photo-detectors, a curve showing the variation of these quantities with wavelength is also required. Neither of these quantities gives an adequate description of the performance of a detector, for the first is only useful when one is working with large signals (i.e. well away from the ultimate limit of sensitivity of the detector) and the second only gives a true measure of the limiting sensitivity if the conditions are clearly stated under which this minimum power was determined. The reason is that every receiver consists essentially of a detector unit for converting infra-red radiation into some other form of energy, together with an amplifying device and a display unit, such as a cathode-ray oscilloscope, a meter or a recorder. By increasing the amplification, a small signal from the detector can be increased to any desired size and the limit of sensitivity is finally determined when the fluctuations due to "noise" become greater than the signal. Although it is possible to observe a signal which is less than noise against the background of that noise, it is customary to quote for the sensitivity of a detector, that radiant flux density (R.F.D.) or radiant flux which produces a signal equal to noise at the display unit. The latter is sometimes known as equivalent noise input or E.N.I.

Noise arises both in the detector unit and in the amplifier, although the latter is normally arranged to be less than the former. If, as often happens, the radiation is periodically interrupted before falling on the detector the noise may also be a function of the chopping frequency. Moreover, under these circumstances, the amplifier is usually tuned to the chopping frequency, so that the band-width of the amplifier or the time of response of the display unit will also control the noise level and hence the E.N.I.

In view of these considerations, we propose to use the term *responsivity* for the output of a detector per microwatt of incident radiation. For signals near the limit of detection the best measure of sensitivity is the *equivalent noise input* (E.N.I.). Relevant details of the associated amplifier and display unit should be given and, where possible, a curve showing output for various chopping frequencies, or a figure for the time of response.

We shall now consider the various detectors following the order given above, except that photoconductive and photovoltaic types will be treated together in the order thallium sulphide, lead sulphide, lead selenide and lead telluride,

this being the order in which their sensitivity extends to longer wavelengths in the infra-red.

(ii) *Photographic*

There have been no significant advances in the technique of photographic detection of infra-red radiation comparable with those achieved in other methods and we do not propose to consider this method further in this report.

(iii) *Phosphors* *

The detection of infra-red radiation by utilizing its property of stimulating phosphorescence has been investigated to a considerable extent recently, but this method has had few applications since (*a*) it is much less sensitive than most other methods, (*b*) it is limited to wavelengths below 1·5μ, and (*c*) it requires re-activation of the phosphor by visible or ultra-violet light. This method has, however, the great advantage that, like the photo-emissive method, it allows one to obtain a direct infra-red "picture" or image. It has been known for a long time that the phosphorescence of certain materials can be enormously enhanced by allowing infra-red radiation (wavelength about 1μ) to fall on them. The mechanism of this process is still very uncertain but much experimental work was carried out in U.S.A. during the war, which has considerably clarified the conditions favouring it, while some progress has been made on the theoretical side both in this country and in U.S.A. during the same period.

When a phosphorescent material is being excited by visible or ultra-violet light, the visible radiation which is emitted is termed fluorescence. After the exciting source has been extinguished the emitted radiation is known as phosphorescence or afterglow. If the phosphorescent material belongs to the class known as "*impurity activated*" then Urbach and others (1946) showed that infra-red radiation will stimulate the phosphorescence more effectively if a second or "*auxiliary activator*" is present. Thus, Urbach and his co-workers (1946) have shown that alkaline earth sulphides (such as calcium and strontium) can have as dominant activator, traces of cerium, manganese or europium, and for such impurity activated phosphors, samarium is a suitable auxiliary activator. The emission spectrum of the phosphor is determined by the dominant activator, the maximum moving from 0·485μ to 0·58μ as we proceed from cerium to europium. The sensitivity to infra-red stimulation is conditioned solely by the auxiliary activator. The maximum response in the case of samarium is at 1·03μ.

These alkaline earth phosphors are more efficient than those of the zinc and cadmium sulphide class which were largely investigated by Fonda (1946) and Miller and Barnet (1946). In this case lead or copper can be used as dominant activator with copper as auxiliary activator. It should be added that phosphors of this class seem to show more promise as detectors of infra-red of rather longer wavelengths, especially when used at low temperatures.

The interpretation of these results is not yet fully understood. However, studies of the decay times and of the temperature dependence of these phenomena have enabled a rough picture to be made in terms of the conventional quantum

* We are indebted to Dr. G. F. J. Garlick for allowing us to see the manuscript of his forthcoming book *Luminescent Solids* (Oxford University Press) in preparing this section.

mechanical theory of bands of energy levels in solids. It is generally accepted now that in a phosphorescent solid the energy levels may be schematically represented as in figure 1, the lowest band being completely full and the conduction band empty of electrons in the case of an insulator. When the prim ary activator is added new levels ("luminescent centres") are created, localized near the impurity and occupied by electrons. Fluorescence phenomena are due to the excitation of these electrons in the impurity centres by the absorption of the exciting radiation to higher levels from which they return (often in stages), the energy liberated on their return appearing as the fluorescent radiation. To explain the phenomenon of phosphorescence it is further postulated that below the conduction band there exist thermally metastable levels or "*electron traps*". To eject an electron from these requires a small amount of energy (equal to the depth of the trap) of the order of 0·1 electron volt. This may be provided by ordinary thermal exchanges in the crystal, and from a study of the temperature dependence of the phosphorescence Randall and Wilkins (1945) have attempted to deduce the distribution of these electron traps. The energy required to liberate the electron from the trap may also be provided by direct absorption of the appropriate infra-red radiation. This would explain the effect of infra-red radiation in stimulating phosphorescence. If this explanation were correct then the infra-red absorption spectra of phosphors might be expected to show a close correlation with the electron trap distributions as determined by experiments on thermo-luminescence. Although similarity exists in some cases, there are many phosphors for which this is not the case. The mechanism of the decay of phosphorescence is certainly more complex than we have indicated above in that electrons ejected from the traps may be absorbed directly into empty luminescence centres with or without emission of radiation. From the work of Randall and Wilkins (1945), Garlick and Wilkins (1945) and Garlick and Gibson (1947) it is clear that such recombination with empty luminescence centres can be a very important factor. Furthermore, the explanation of the quenching, as opposed to the stimulation, of phosphorescence by infra-red radiation is not at all clear on the above model.

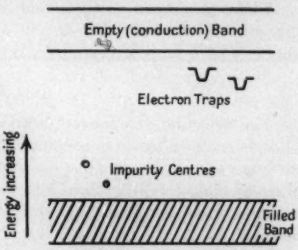


Figure 1. Distribution of electronic levels in a phosphorescent solid.

Much more work requires to be done on this subject, the complexities of which have only been hinted at in this outline. In particular, it would seem most important to conduct studies at very low temperatures if there is to be any hope of extending this method of detecting infra-red radiation to longer wavelengths. In this connection mention may be made of the quenching of the phosphorescence of a zinc cadmium sulphide phosphor (activated by copper) by infra-red radiation of wavelength as long as 2·7μ when working at the temperature of liquid air (Daly and Sutherland, unpublished).

(iv) *Photo-emissive image converters*

The image converter is an instrument which transforms an invisible infra-red image into a visible one. It consists essentially of an objective lens or mirror, an image tube operated from a suitable high voltage source and an eyepiece or viewing lens. Such instruments are not new; indeed many designs existed before 1939 (Myers 1939) but it is probably fair to say that these were little more than laboratory curiosities. During the recent war, considerable effort was concentrated in Britain, America and Germany on their development for military purposes. The Services demanded high sensitivity, good resolution, ruggedness and easy maintenance. For some applications a lightweight "pocket size" instrument was required; on the other hand, equipment was built in Germany which incorporated complex 24″ diameter lens mirror systems and weighed many hundredweight. The American equipment has been described in a number of publications (Morton and Flory 1946) and German instruments (B.I.O.S. 1946) were shown at an Exhibition of Electronic Equipment held at Earls Court, London, in March 1946 (*Electronic Engineering* 1946). A paper on British service equipment, known as "R.G. receivers" (*A.R.L. Reports* E.300, 1939–46), is awaiting publication.

The infra-red image tube consists of a sensitized cathode and fluorescent anode suitably arranged in an evacuated glass envelope, so that photoelectrons, liberated from the cathode by infra-red radiation, are accelerated on to the anode to produce a visible fluorescent picture. All standard tubes use semi-transparent cathodes of the Cs–O–Ag type, and anodes usually of willemite; it is indeed remarkable that a vast amount of research in each country has not succeeded in producing superior alternatives to these well-known materials. The spectral sensitivity of these cathodes is well-known (Sommer 1946); there is a peak at $0{\cdot}8\mu$ and the long-wave threshold lies near $1{\cdot}3\mu$. The sensitivity of a well activated cathode is about 20–40 microamps per lumen for unfiltered light from a tungsten filament lamp at a colour temperature of 2870° K. (Morton and Flory 1946).

The various designs manufactured during the war differed chiefly in the means used to produce an electron image on the anode. Two types, the standard German tube made by A.E.G. and the British "R.G." tube developed by E.M.I. Ltd., are shown in figure 2. The former was developed in the first place as a picture-forming instrument whereas the latter was initially designed for viewing point sources in signalling applications; eventually both tubes were employed in each of these spheres. The German tube incorporates electron-optical focusing, so that electrons leaving a point on the cathode in slightly diverging directions are brought to a single corresponding point on the anode; in the R.G. tube, which is based on a design by Holst and others (1934), the cathode and anode are plane and parallel, the electrons being drawn across the 5 mm. gap between them, with very little spreading because of the high value of the electrostatic field (Pratt 1947).

It is interesting to compare these two image tubes. The non-focusing type is more compact and lends itself much more readily to mass production, two factors which are very important in war. The image is free from distortion,

and is equally sharp over the whole field. There are no critical potentials. Moreover, the current required is extremely small: a bright picture is obtained with about 10^{-8} amp., and weak distant infra-red beacons are visible with only 10^{-12} amp. flowing, so that simple high voltage sources of the voltaic pile type are sufficient for many purposes. The focusing type of image tube, on the other hand, has some fundamental advantages. In the first place the accelerating voltage between cathode and anode can be made greater in this type of tube (and consequently a brighter image can be obtained) without introducing troublesome field emission from the cathode or elsewhere due to high field strength. With weak images this increased brightness is very important, as the detail which the eye can then perceive is a function of the brightness itself. Secondly, electron-optical reduction is possible when focusing is employed; this again increases the brightness of the image. Thirdly, electron-optical inversion simplifies the associated optical systems. The image produced on the cathode by the objective lens is an inverted one, so that the fluorescent image is upright and can be viewed through a simple wide-field eyepiece. For many applications, however, for example where point sources of infra-red radiation are viewed as in signalling and navigation, an inverted image is not a very great drawback, so that the R.G. type of tube combined with a simple plano-convex eye-lens can still be used. An example of this is the "Type U" receiver, shown in figure 3.

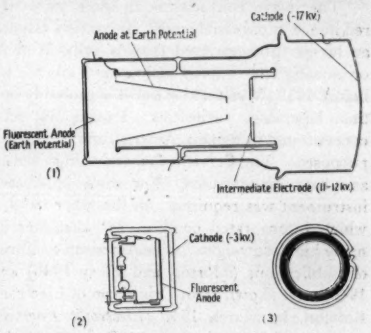


Figure 2. Image converter tubes (German—A.E.G., and British—R.G.)

(1) The simplified A.E.G. tube introduced in 1943. Focusing type.

(2) R.G. tube (side view) } 1941. Non-focusing type.
(3) R.G. tube (end view) }

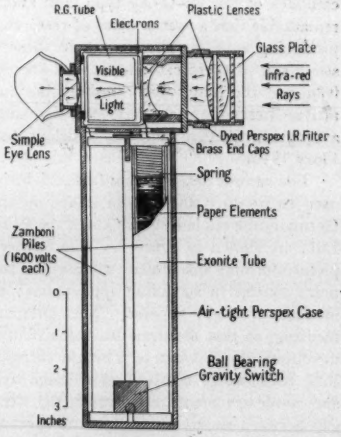


Figure 3. British image converter type BR/U3 (R.G. receiver).

A successful technique has been devised in Germany for depositing aluminium layers, about 0·5 to 1 micron in thickness, over the fluorescent material of the anode. The effect of this is threefold. The good electrical conduction which it affords across the anode eliminates any chance of the latter charging up to produce retarding potentials. Furthermore, with accelerating voltages above 6 kv., a brighter image is obtained, since the retardation of the electrons in the metal layer is more than compensated for by the reflection, towards the observer, of light which otherwise would be directed towards the cathode. Finally, since no visible radiation from the anode falls on the cathode any "fogging" effect due to this is eliminated.

When an image converter is to be used for detecting distant point sources of infra-red radiation, its performance can be usefully defined by the candle power of a suitably chosen standard source, which can just be seen through a perfectly clear atmosphere at a distance of one mile. The unit is the "mile candle" and obviously high sensitivity is associated with low "mile candle" values. A tungsten filament lamp, with an agreed colour temperature and screened by a standard infra-red filter is normally used as a standard source. In this country, two standard filters, "S.S. Red" and "S.S. Black" have been used for this purpose. The "Red" filter is almost identical with the well-known Wratten 87, and the "Black" filter, as the name indicates, is very much more opaque to visible light. With an average image tube the ratio of "Red" to "Black" sensitivities is of the order of 5 to 1. Some idea of the sensitivity normally achieved can be gauged from the fact that the acceptance test of the Type U instrument shown in figure 3 was 5 "Red" mile candles and 25 "Black" mile candles. In practice observers vary enormously in their ability to see faint point images in an image converter. Also an observer's ability to detect the image depends on his degree of dark adaptation, his comfort, and a number of other factors. In order, therefore, to obtain uniform results, comparative methods are necessary, and artificial receivers showing a constant standard signal resembling that from a distant point source are used for comparison purposes during tests.

For viewing extended objects, two tube characteristics are important, namely image brightness and resolution. As has been previously mentioned, in most applications the level of image brightness is such that the eye itself limits the amount of detail which can be appreciated. At medium brightnesses, 400 black and white lines per inch on the cathode of an R.G. tube can be resolved. For tubes with electron focusing the definition at the centre of the field of view is usually slightly better than this, but falls to a lower value away from the centre.

(v) *Photoconductive and photovoltaic cells*

(a) *Thallium sulphide photoconductive cells.*—The thallium sulphide infra-red sensitive photoconductive cell was first described by Case (1920) and since then a number of papers describing its properties and methods of preparation have appeared (e.g. Majorana and Todesco 1928). During the war this cell received much attention in military research laboratories in Britain, America, Germany, Italy and probably elsewhere. Publication of this work is far from complete but some information is available (Cashman 1946, Starkiewicz 1944, Hewlett 1947).

Recent work has been largely concerned with improving sensitivity, reproducibility in manufacture, and stability to high light intensities. Lack of stability characterized all pre-war thallium sulphide cells, rendering them unsuitable for many potential applications. Reports on the processes used during manufacture to overcome this and other difficulties will no doubt be published soon. A typical spectral sensitivity curve is shown in figure 4. The long wave threshold

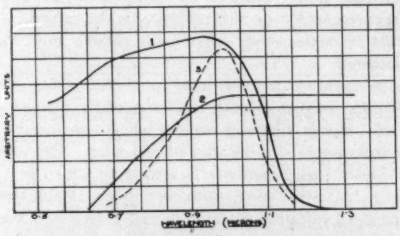


Figure 4. Properties of thallium sulphide photocells.
Curve 1. Spectral sensitivity of thallium sulphide photoconductive cell.
Curve 2. Transmission of thin layer of thallium sulphide.
Curve 3. Spectral sensitivity of thallium sulphide photovoltaic cell.

is at 1·4 μ and the peak response in the neighbourhood of 1 μ. Good sensitivity continues into the visible region. The specific resistance of sensitive thallium sulphide is high, so that grid shaped electrodes have to be used to reduce the resistance of the cell. A typical cell is shown in figure 5. The thallium sulphide is deposited between the grid electrodes.

Figure 5. A thallium sulphide photo conductive cell.

The resistance of the original Case cell was of the order of 100 megohms, but this inconveniently high value has now been reduced to a few megohms. According to Cashman (1946) the response curves of thallium sulphide cells depart from linearity with radiant intensity above 0·02 ft.-candles, though saturation is not reached even at 10 000 ft.-candles. They show rapidly falling responsivity with increasing frequency. The response of those made at the Admiralty Research Laboratory falls by 70% as the modulation frequency is increased from 50 c.p.s. to 250 c.p.s., and by a further 15 % at 600 c.p.s. No measurements on the frequency/noise characteristics have so far been published. One would expect the behaviour to be similar to lead sulphide which has been very carefully studied in this respect, i.e. a falling noise characteristic with increasing frequency; this is confirmed by Cashman's remark (1946) that the signal/noise ratio remains practically constant for modulation frequencies not exceeding a few thousand cycles per second.

(b) *Thallium sulphide photovoltaic cells.*—Kolomiez (1938) and Nix and Treptow (1939) have described the manufacture and properties of thallium sulphide photovoltaic cells of the so-called barrier-layer type. The latter authors

described their cells as laboratory apparatus not available for commercial distribution, probably because of ageing effects which had not been eliminated.

The cells are made by exposing an evaporated layer of thallium supported on a copper or nickel disc, to a glow discharge in H_2S gas or sulphur vapour. The sulphide layer is then coated with a thin film of sputtered platinum or gold. No mention is made of the use of oxygen. This is rather strange as the presence of oxygen is a fundamental requirement for sensitive photoconductive cells using either lead sulphide or thallium sulphide.

A typical spectral response curve taken from Nix and Treptow is shown in figure 4. The maximum is at 0·98 μ, while long and short wave thresholds are in the neighbourhood of 1·5 μ and 0·65 μ, respectively. Curves for the thallium sulphide photoconductive cell and the transmission curve for a thin layer of thallous sulphide are also given by these authors, who suggest that visible radiation is filtered out by the thallium sulphide before the sensitive region of the cell is reached.

Nix and Treptow give a figure of 6000 microamps per lumen for the best responsivity obtained to light from a tungsten filament lamp at a colour temperature of 2848° K. At first sight this may appear high compared with the 450 microamps per lumen often quoted for the selenium photovoltaic cell, but it must be remembered that the spectral emission curve of the lamps greatly favours the thallium sulphide cell. The less sensitive cells of this type exhibit accurate proportionality between either short circuit current or open circuit voltage and incident light intensity up to at least 1 lumen. The more sensitive cells, however, show saturation effects at lower intensities.

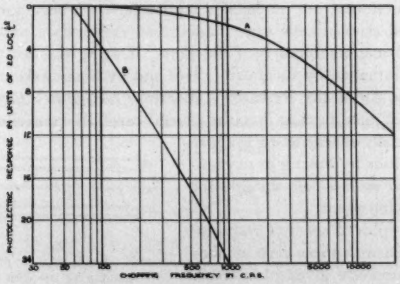


Figure 6. Frequency response characteristics of thallium sulphide photovoltaic cell.
A. Thallium sulphide photovoltaic cell.
B. Selenium photovoltaic cell. From Nix and Treptow.

The frequency response characteristics of thallium sulphide and of selenium photovoltaic cells are shown in figure 6. Even at 20 000 c.p.s. the response of the former is still 25 % of the maximum.

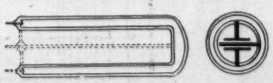
(c) *Lead sulphide photoconductive cells.*—The detectors of infra-red radiation in common scientific use prior to the war consisted of photocells and photographic plates, sensitive between the visible region and about 1·2 μ, and thermodetectors such as thermocouples or bolometers for longer wavelengths.

Considerable improvements in these detectors have taken place during the last few years, but the development which is probably of most interest in the infra-red field is the lead sulphide cell which has extended photo-detection from 1·2 μ to about 3·5 μ. Lead selenide and lead telluride cells have also been made which are sensitive to about 5μ and 7μ respectively, but only lead sulphide cells have yet been manufactured in numbers.

Credit for the initial development of these cells goes to the Germans, who concentrated considerable effort on the problem prior to and during the war because of their obvious military application for detecting the hot funnels of ships and the glowing exhaust pipes of aircraft. Field equipment based on them was never used successfully but the cells themselves were made in considerable numbers. Towards the end of the war the development of lead sulphide cells was taken up in this country and America, but very little of this recent work has yet been published.

Photoconductive layers have been made in two distinct ways which will be very briefly described. The first is the method of chemical deposition. These are often referred to as ELAC cells and Dr. Kutschar of Electroakustik A.G., Kiel, Dr. Kaspar of A.E.G. and the firm Kast and Ehinger, seem to have been mainly responsible for their development (Oxley 1946). The photosensitive layers are obtained by deposition from an aqueous solution of thio-urea by the action of lead acetate and sodium hydroxide solutions. After "tempering" and degassing, the layers are sealed into an evacuated glass envelope. Since 1944, cells of this type have been made at the Admiralty Research Laboratory (*Admiralty Reports* 1939–46 c) and more recently by British Thomson-Houston Research Laboratories, but no detailed reports have yet been published.

The second class of cells were made by an evaporation process by Gudden and Treu and manufactured by Zeiss-Ikon. During the last three years they have been investigated by Cashman (1946) and by Sosnowski and others (1947) working at the Admiralty Research Laboratory (*Admiralty Reports* 1939–46 c). The method of manufacture involves several carefully controlled processes but consists essentially of sensitizing evaporated lead sulphide by baking in oxygen. Figure 7 shows such a cell suitable for use with a cooling agent.



(*a*) (*b*)

Figure 7. Lead sulphide photoconductive cell suitable for use with refrigerant.
(*a*) Section.
(*b*) End view, showing three electrodes.

Some examples of spectral response curves obtained with these cells at normal temperatures are given in figure 8. The shape of these curves and the positions of peak response vary considerably from cell to cell, depending on the method of manufacture. According to Moss (1947) the wavelength corresponding to peak response and the long-wave threshold move progressively towards longer wavelengths as the temperature of the layer is lowered.

When assigning a sensitivity to a cell, it is necessary to state the current through the cell, the chopping frequency, the band width of the amplifier and the temperature of the cell, as each of these factors influences the signal to noise ratio. When work on lead sulphide cells began in this country, a standard test

equipment was devised, so that sensitivity measurements made in various laboratories could be directly compared. Briefly, the equipment consisted of an optical bench on which a standard black body and the cell are mounted about 10 cm. apart. The black body, which is 3 mm. in diameter, is normally maintained at 200° c. and the radiation is chopped at 800 c.p.s. before falling on the cell. A tuned amplifier (band-width 30 c.p.s. to 0·707 of peak gain) is employed, which terminates in a cathode follower stage, feeding a rectifier meter, so that r.m.s. noise from the cell is indicated as a steady deflection. The optimum cell current can soon be ascertained, but in the normal test only 800 c.p.s. (which may not be the optimum frequency) is used, and measurements are made with the cell (*a*) at room temperature and (*b*) at the temperature of solid carbon dioxide. The signal/noise ratios obtained with this equipment correspond to a response time of approximately 0·03 seconds.

The total energy of all wavelengths falling on 1 sq. cm. in the neighbourhood of the cell is approximately

$$E = 5600/R^2 \text{ microwatts},$$

where R is the distance in cm. between the cell and the black body. The energy in the region 1 to 3·5μ is roughly 2·4 % of this value.

The order of sensitivity of uncooled cells can be found by reference to the spectral sensitivity curves in figure 8. Here the ordinate on the right gives the

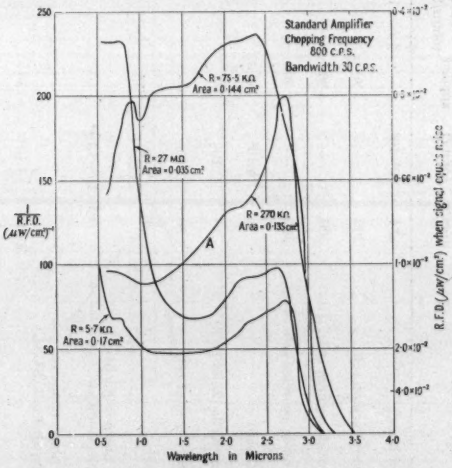


Figure 8. Spectral sensitivity of typical lead sulphide cells.

value of the radiant flux density (R.F.D.) which produces a signal equal to noise. Some sensitivity measurements, for lead sulphide and other types of photocells, made under standard conditions but using a high temperature source, have been given by Allekotte in a German military report. His table of results is reproduced in table 1 (see next page).

Table 1. Sensitivities of German cells (Allekotte 1942)

Type of cell	Area (mm^2)	Resistance (ohms)	Voltage applied	Matching resistance (ohms/800 c.p.s.)	Threshold (lumens)	Threshold (watts)	Amplification required for 1 volt per micro-lumen
Selenium barrier layer	314	10^3–10^5	—	100	2.10^{-6}	$1{\cdot}6.10^{-7}$	3.10^5
Caesium–antimony vacuum cell	300	—	90	10^6	7.10^{-7}	5.10^{-8}	2.10^5
Caesium silver oxide vacuum cell	300	—	90	10^6	1.10^{-6}–5.10^{-7}	8.10^{-8}–4.10^{-8}	2.10^5–1.10^5
Caesium silver oxide gas-filled cell	300	—	90	10^6	2.10^{-7}	$1{\cdot}6.10^{-8}$	4.10^4
Thallofide cell (Zeiss)	1	4.10^7	30	10^6	4.10^{-7}–8.10^{-8}	3.10^{-8}–6.10^{-9}	$1{\cdot}2.10^4$–$3{\cdot}6.10^3$
Lead sulphide netzzelle (Zeiss Ikon)	50	10	—	10	$1{\cdot}5.10^{-6}$	$1{\cdot}8.10^{-7}$	5.10^5
Lead sulphide barrier layer cell	50	10	—	10	4.10^{-5}	3.10^{-6}	3.10^6
(Siemens)	1	10	—	10	2.10^{-5}	$1{\cdot}5.10^{-6}$	
Lead sulphide photoconductive evaporated cell (Gudden)	0·5	10^5–5.10^5	10	10^6	7.10^{-8}–2.10^{-8}	5.10^{-9}–$1{\cdot}6.10^{-9}$	$3{\cdot}5.10^2$–$1{\cdot}7.10^2$
Lead sulphide photoconductive chemically prepared cell (ELAC)							
Uncooled	30–100	4.10^5	50	10^6	1.10^{-7}	8.10^{-8}	$1{\cdot}2.10^4$
Cooled with solid CO_2	30–100	—	150	10^6	2.10^{-8}–1.10^{-8}	$1{\cdot}6.10^{-8}$–8.10^{-10}	$1{\cdot}5.10^3$–5.10^2

Source : Tungsten filament lamp at 2400° K.
Chopping frequency : 800 cycles per second.
Amplifier : Flat response 30–10 000 cycles per second.
Low resistance cells used with a transformer with stepped primary winding.

The equivalent noise input of a cell will be small (i.e. the sensitivity will be high) if the responsivity measured in microvolts per microwatt of incident energy is high and the noise from the cell is low. It is, therefore, important to know

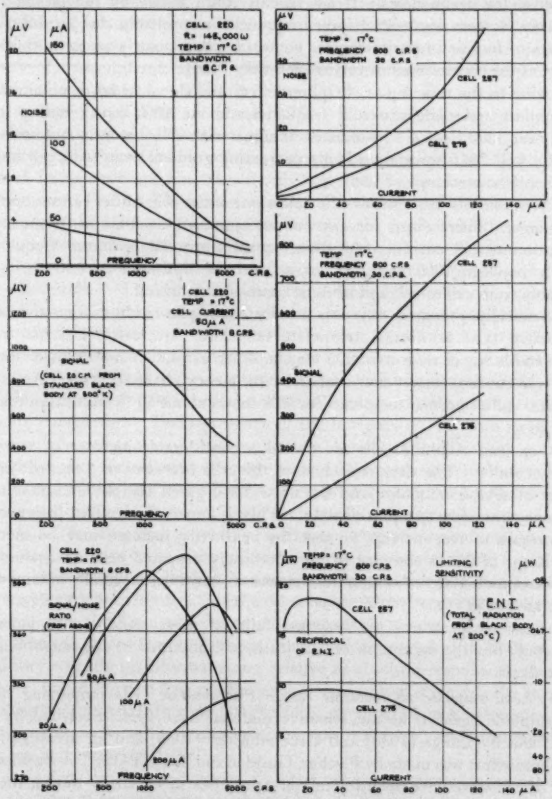


Figure 9. Signal, noise and sensitivity characteristics for typical A.R.L. evaporated lead sulphide cells.

how these factors vary with cell current, chopping frequency and cell temperature.

Figure 9 shows noise, signal, and signal to noise ratio graphed against cell current and frequency for some evaporated A.R.L. cells at normal temperatures.*

* The curves showing variation of signal and noise with frequency were obtained at the Telecommunications Research Establishment, Malvern; we are grateful for permission to reproduce these before publication.

It will be observed that besides the thermal agitation noise present in the cell with zero current flowing (which is independent of frequency), there is a large noise component which depends upon the value of the current and which is greater at low frequencies. It has recently been shown by Macfarlane (1947) that this increase in noise at low frequencies is probably due to diffusion of clusters of mobile impurity centres on to the contact surfaces as in Schottky's theory of the flicker effect in valves (Schottky 1926).

Owing to the time lag of the photo-effect, the signal which is obtained for a given illumination decreases at frequencies above 1000 c.p.s., falling to half value near 3000 c.p.s. The increase of signal with cell current is not necessarily linear ; cell 257 shows a gradually decreasing gradient even with cell currents less than 50 microamps.

The sensitivity curves show the behaviour of the cells under various operating conditions. There exists for any cell an optimum modulation frequency and an optimum cell current. At room temperature the optimum frequency is usually between 1000 and 2000 c.p.s., but the optimum current varies considerably from cell to cell and in some cases is very critical.

Chemically prepared cells are usually about one-tenth as sensitive as the evaporated type at normal temperatures. They are less noisy, but have a considerably lower responsivity. On the other hand, they have a lower response time, and consequently better sensitivity/frequency characteristics. An average chemical cell at normal temperatures still shows about 50 % maximum response at 10 kc/s.

When lead sulphide cells are cooled a considerable increase of sensitivity usually results. The dark resistance of the cells increases in a manner characteristic of semi-conductors, and the noise for a given current and modulation frequency increases proportionately. This is more than offset, however, by the increase in responsivity, so that the sensitivity increase may be increased by a factor of 20, in the case of the chemically prepared cells on cooling with solid carbon dioxide. A four-fold increase of sensitivity is usually obtained with evaporated cells.

At low temperatures the response falls off more rapidly with increasing frequency, but the maximum sensitivity is still obtained in the neighbourhood of 1 to 2kc/s.

(d) *Lead sulphide photovoltaic cells.*—Photovoltaic cells employing natural lead sulphide (galena) as the sensitive material have been described by Bose (1904) and by Lange (1930) and Gruetzmacher (1930) and an investigation of this photo-effect was made by Fischer, Gudden and Treu (1938). Two improved types of cell were developed for military purposes in Germany during the war. One of these the Siemens cell, was made by evaporating a thin layer of silver on to a tablet of synthetic lead sulphide. It was less sensitive than other types of lead sulphide cell and does not appear to have been applied in any way.

The second type of photovoltaic cell, however, was more successful. It was known as the gauze cell (netzzelle) and consists essentially of a polished tablet of Sardinian galena, suitably treated, against which is pressed a gauze of very fine molybdenum or tungsten wire. When the cell is illuminated through the gauze electrons flow to the latter from the crystal.

Allekotte has given a figure of $1{\cdot}8\times10^{-7}$ watts as the minimum detectable power (E.N.I.) of these cells using as source a tungsten filament lamp at 2400° K. (table 1). The radiation was chopped at 800 c.p.s. and a wide-band amplifier (30–10 000 c.p.s.) was used. The galena gauze cell is superior to the photoconductive cell in one respect, which is important in daylight use; it retains its responsivity to modulated light and its low noise even in the presence of a high intensity of unmodulated light. With 20 000 lux of additional radiation the response of a gauze cell only falls by about 20%, while that of an evaporated photoconductive cell falls by about 98%. The spectral response appears to be similar to that of lead sulphide photoconductive cells.

A new type of photovoltaic cell has recently been developed at the Admiralty Research Laboratory and has been briefly described by Starkiewicz, Sosnowski and Simpson (1946). The sensitive material is lead sulphide but it is very probable that the method of manufacture can also be applied to other materials. The most interesting feature of these new cells is their high resistance and high voltage response, as against the low resistance and high current response of normal barrier layer cells. The cells consist of a layer of suitably sensitized lead sulphide about 1 micron thick, on glass, between graphite electrodes. Photovoltaic and rectifying properties are developed in the layer by passing a current through it at a temperature sufficiently high to ensure mobility of the oxygen ions (about 250° C.). The spectral response curve is the normal one characteristic of lead sulphide. The photo-E.M.F. is of the order of 2 volts with cells having resistances of about 2 megohms. The sensitivity of the few cells of this type so far made is, however, less than that of photoconductive lead sulphide cells. The minimum detectable power for radiation between 1 μ and 3·5 μ is about 4.10^{-8} watts.

(e) *Lead selenide and lead telluride.*—The photoconductive properties of lead selenide and telluride have not been so carefully investigated as those of the sulphide. The important property of cells employing photoconductivity in these substances is the extension of the spectral response curve further into the infra-red region. According to Kutscher peak response for lead selenide is at 3 μ and the long wavelength limit is 4·5 μ. On the other hand, Allekotte gives a curve showing the long wave threshold at 5·5 μ which agrees with results obtained by Treu. Lead telluride, according to Krentzien, has its long wavelength limit at 5·8 μ. Until very recently no information on relative or absolute responsivity has been available, but in the past few months very promising lead selenide cells have been made in this country (Simpson 1947) which are superior to a Schwarz vacuum thermocouple as far as 4 μ, with sensitivity maxima at 1·6 μ and 3·3 μ (Blackwell, Simpson and Sutherland 1947).

Lead selenide cells were prepared by Treu, using an evaporation process which, as in the case of the sulphide, involves a bake in oxygen. Simpson (1947) also used an evaporation method but no bake in oxygen. The chemical process for manufacturing lead sulphide cells was also used to produce some selenide cells by precipitation from a mixture of lead acetate and seleno-urea. Full details are not available but it is known that an initial layer of lead sulphide was formed to act as a "seed" for the deposition of the selenide.

Krentzien used an evaporation process for experimental investigation of the photoconductive properties of lead telluride. The sensitive layer was produced by

heating together lead and tellurium in a graphite crucible using an oxygen or nitrogen atmosphere. After purification by sublimation *in vacuo* the telluride was evaporated on to an electrode plate in the usual way and baked in oxygen at various pressures and temperatures. The cells thus produced varied considerably in behaviour. They usually showed no sensitivity at normal temperatures, a little at −78° c., but improved considerably at liquid air temperatures. Others showed sensitivity at liquid air temperature but no sensitivity at the temperature of solid carbon dioxide.

[*Note added in proof.*—Since the above was written, very successful lead telluride cells have been produced in this country by the evaporation method (Blackwell, Simpson and Sutherland 1948). These cells are comparable in sensitivity at their point of maximum response (2·2 μ) with the best lead sulphide cells and are superior in performance to a Hilger-Schwarz vacuum thermocouple as far as 4·6 μ; the cells still show some response at 5·5 μ. All observations were made with the lead telluride at liquid air temperature as these cells showed no photo-sensitivity at room temperature.]

Further details concerning all the new infra-red photocells have just been published by Elliott (1947).

(f) Recent work at the Admiralty Research Laboratory (*A.R.L. Reports* 1947), of which a very brief summary has been published (Sosnowski, Starkiewicz and Simpson 1947, Sosnowski, Soole and Starkiewicz 1947), has thrown light on the mechanism of photoconductive and photovoltaic effects in sensitized layers of lead sulphide and other materials, and a theory of the effects has been proposed. According to these authors a layer of sensitized lead sulphide consists of a large number of microcrystals, some of which may contain excess of lead and others excess of oxygen as impurity centres. The boundaries between such crystals will have rectifying properties and control the conductivity of the layer.

When such a layer is illuminated with radiation of suitable wavelength two effects are produced. Firstly the resistance of the barriers is reduced because of the ejection of electrons into the conduction band (the photoconductive effect), and secondly, because of the diffusion of electrons across the barrier from the conduction band in the semi-conductor containing oxygen impurity, a photo-E.M.F. is generated (the photovoltaic effect). If a layer contains many barriers distributed at random, the photovoltaic and rectifying effects will not be observable, but when, by applying the treatment described in (d) above, the small rectifiers are aligned, both rectifying and photovoltaic effects are observed. The presence of these small photovoltaic elements in ordinary photoconductive layers which, as a whole, show no photovoltaic effect, has been beautifully demonstrated by moving a very small spot of light across such a layer, when the small voltages associated with randomly oriented barriers were observed.

(vi) *Thermocouples*

The thermocouple is the traditional tool for the detection of infra-red radiation and until 1939 single junction vacuum thermocouples were almost universally used for precision work, e.g. in infra-red spectroscopy. Probably the best designs were those due to Pfund (1937), Firestone (1930) and Cartwright (1934),

which have been described very fully by Strong (1939). In such thermocouples the thermoelectric elements consisted of thin wires of a bismuth/tin and a bismuth/antimony alloy; the receiver was a very thin strip of silver or gold foil "blackened" with a lamp or metallic black. The responsivity of a thermocouple of this type is about 15 microvolts per 100 microwatts/cm². The time of response (viz. about 3 seconds) is a serious drawback for some purposes, but tolerable in spectroscopic work where speed is a secondary consideration compared with resolving power (e.g. work with a diffraction grating on the rotational fine structure of absorption bands).

Early in the course of the war, the need arose for a sensitive thermocouple which was also sufficiently rapid in response to enable spectra to be recorded quickly for analytical purposes. This was fulfilled admirably by the Schwarz thermocouple manufactured by Adam Hilger Ltd., which represented a considerable advance in performance on any previous thermocouple. It appears to be more sensitive, and is certainly more rapid and more robust than any of its predecessors. The method of construction is novel, as are the materials used for the sensitive elements. The latter are complex semi-conductor alloys ground into small cones. The receiver consists of a strip of blackened gold foil connecting the tips of the cones. The responsivity is approximately 20 microvolts per 100 microwatts/cm², which is not very different from the Pfund type of thermocouple, but the time of response of the Schwarz thermocouple is less than 0·2 sec., i.e. at least 15 times faster than the Pfund type.

In the same direction of increasing speed without unduly sacrificing sensitivity, Roess and Dacus (1945 a) made a fast thermocouple by condensing thin films of bismuth and antimony on to a film of the plastic Formvar and "blackening" the junction with evaporated gold. In this they were following method originally developed by Harris and his co-workers (Harris 1934, 1946, Harris and Johnson 1933, 1934, Harris and Scholp 1940), but have been more successful in reducing the resistance of the couple, so getting a better signal/noise ratio. The time constant of a typical thermocouple of this type is about 0·15 sec. for 95% of full response. The responsivity is given as 6–7 microvolts per 100 microwatts/cm², the resistance of the thermocouple being about 35 ohms.

The design of fast thermocouples using the conventional construction of short, fine wires of Bi/Sb and Bi/Sn has recently been investigated by Hornig and O'Keefe (1947). They point out that the evaporation technique has serious disadvantages in that the resistance of thin films is higher and the thermoelectric power is usually much lower than for the same materials in bulk. By using extremely short thermoelectric wires (0·2 to 0·5 mm. in length) they have been able to reduce the heat capacity so that times of response are as short as for couples made by the evaporation process. The responsivity of one such thermocouple is given as 3·25 microvolts per 100 microwatts/cm², the time constant being 0·2 second for 88% of full response and the resistance 5 ohms.

It will be noticed that so far we have only quoted figures for responsivity. Before endeavouring to compare the equivalent noise input for the various thermocouples, it will be well to discuss in some more detail the factors governing the ultimate sensitivity of a thermocouple.

A thermocouple, being one of the class of thermo-detectors, must be limited in

its ultimate performance by the natural fluctuations in temperature which occur in the detector element. It is somewhat surprising that this fundamental limitation has only recently been fully investigated. Previously the whole receiver had had been considered, i.e. including the leads and the galvanometer. These introduce additional sources of noise, among other complications, and it is well therefore to start with an idealized thermocouple in which temperature fluctuations in the receiver element are the only source of noise. It has now been shown independently by Daunt (1945 a), Lewis (1947) and Milatz and van der Welden (1943) that the minimum power detectable as a sinusoidal variation of frequency by an isolated body of area A in which the only thermal losses are by radiation is given by

$$W^{\text{min.}}_{\text{thermal}} = 4\sqrt{(A\sigma \boldsymbol{k} T^5 \Delta f)}, \qquad \ldots\ldots(1)$$

where σ is Stefan's constant, $\boldsymbol{k}$ is Boltzmann's constant, T the absolute temperature and Δf the band-width. Putting in numerical values we get

$$W^{\text{min.}}_{\text{thermal}} = 5{\cdot}7 \times 10^{-12} \text{ watt}$$

for $A = 1$ mm², $T = 300°$ K. and $\Delta f = 1$ c.p.s. This is the absolute limit which no actual thermo-detector can attain since heat losses by conduction can be shown to increase this value.

However, this is not the only factor controlling the theoretical performance of a thermocouple. In practice it is much more likely to be the Johnson noise arising from fluctuations due to Brownian motion of electrons in the circuit. The expression for the voltage fluctuation due to this is well known to be (Johnson 1928)

$$E_J(\text{r.m.s.}) = \sqrt{(4\boldsymbol{k}TR\Delta f)}, \qquad \ldots\ldots(2)$$

where R is the total circuit resistance. Putting in the numerical values we get

$$E_J(\text{r.m.s.}) = 1{\cdot}3 \times 10^{-10} \sqrt{(R\Delta f)} \text{ volts},$$

where R is in ohms. From the figures given above for various thermocouples, it is seen that an average figure for the responsivity (ρ) of a modern thermocouple is about 0·1 microvolt per microwatt/cm². The minimum detectable power as determined by Johnson noise is therefore given by

$$W^{\text{min.}}_J = \frac{1{\cdot}3 \times 10^{-10} A\sqrt{(R\Delta f)}}{\rho} \text{ watts}, \qquad \ldots\ldots(3)$$

or substituting $\rho = 0{\cdot}1$, $A = 1$ mm², $R = 10$ ohms and $\Delta f = 1$ c.p.s.

$$W^{\text{min.}}_J = 4{\cdot}1 \times 10^{-11} \text{ watt}.$$

It will be observed that this is more than seven times higher than the limit just calculated, due to thermal fluctuations, and in general it is Johnson noise which limits the sensitivity of thermocouples.

How closely this limit is approached in practice it is a little difficult to say, as sufficient data are not always given, particularly with regard to the band-width employed. Hornig and O'Keefe (1946), however, claim to have come very close to it. The relevant figures for one of their best thermocouples and other typical ones are collected in table 2, in which the fifth column gives the minimum detectable power as limited by Johnson noise. We may accept this as a safe lower limit for the equivalent noise input for these thermocouples for a band-width of 1 c.p.s. Unfortunately Roess and Dacus do not give the area of their detector element. In column 6 we have calculated the quantity $\rho R^{-\frac{1}{2}}$. It is clear from equation (3)

that so long as Johnson noise is the controlling factor in the performance of a thermocouple then the sensitivity for a given radiation flux is determined by $\rho R^{-\frac{1}{2}}$.

Table 2. Characteristics of various thermocouples

Maker	Resistance (ohms) R	Responsivity (μ volt/μ watt/cm^2) ρ	Receiver area (cm^2) A	$W_J^{\min.}$ (watts) $\times 10^{-11}$	$\rho R^{-\frac{1}{2}}$	Response time in seconds (approx.)
Hilger–Schwarz	100	0·20	0·004	2·5	0·02	0·15
Weyrich–Pfund	25	0·15	0·016	6·9	0·03	3·0
Roess and Dacus	35	0·065	?	?	0·01	0·14
Hornig and O'Keefe	5	0·0325	0·005	4·4	0·015	0·2

When this quantity is evaluated for the different thermocouples it is interesting that it does not vary by more than a factor of 2 for all the new fast couples.

(vii) *Bolometers*

The bolometer, which had fallen out of favour before the war as inferior in sensitivity to the vacuum thermocouple, has re-appeared in new forms and has displaced the thermocouple where speed of response is the primary consideration. One of the most successful of the new bolometers is the thermistor type made at the Bell Telephone Laboratories by Becker and his colleagues (Becker, Green and Pearson 1946). The semi-conducting material used is a mixture of the oxides of nickel, manganese and cobalt, but, as with the Schwarz thermocouple, the full details of the construction have not yet been published. Bolometers of this type can be made with a time constant of 4–5 milliseconds and a responsivity of 0·8 μvolt/μwatt/cm^2 at an interruption rate of 15 c.p.s., the resistance of the element being about 3×10^6 ohms and the receiver area $2 \times 0{\cdot}2$ mm. The flake of semi-conductor is in this case backed by glass or quartz, which, it is stated, acts as a heat "sink", so increasing the speed. With an unbacked flake in air, a time constant of 25 to 30 milliseconds is obtained, but the responsivity is higher at 2 μ volt/μ watt/cm^2. It is claimed (Brattain and Becker 1946) that the faster variety gives "a signal which is equal to noise when 2×10^{-8} watt power falls on it for 3 milliseconds for 30 c.p.s. band-width"; for signal equal to noise in the slower type, an incident power of 10^{-10} watt is quoted for radiation interrupted at 15 c.p.s. using an amplifier with a band-width of 2 c.p.s. (Becker and Moore 1946). It will be observed that the slower of the B.T.L. bolometers appears to be somewhat higher than the new rapid thermocouples in equivalent noise input but distinctly faster in response.

Other rapid bolometers developed during the war have been described by Aiken, Carter and Phillips (1946) who used a very thin film of gold evaporated on to cellulose nitrate film, and by Billings, Barr and Hyde (1947 a). The latter used nickel, evaporated on to nitrocellulose and obtained a time constant of 4 milliseconds; the equivalent noise input is given as $3{\cdot}3 \times 10^{-8}$ watt for radiation modulated at 30 c.p.s., where the band-width of the system is 100 c.p.s.

An important new type of bolometer has been devised by Andrews and his collaborators (Andrews, Milton and de Sorbo 1946, Milton 1946), in which the element is held in the transition range between normal and super-conductivity. The idea of a super-conducting bolometer must have occurred to many physicists, but the inherent difficulties in working with liquid helium seemed to make it a rather impracticable proposition. The present success is based on the discovery by Horn and others (1942) that columbium nitride becomes super-conducting at a temperature of about 14° K., which is attainable with liquid hydrogen. The time constant of this bolometer reaches the extraordinary figure of 0·5 millisecond for 63% of full response, while the equivalent noise input is given as 5×10^{-10} watt. The band-width is not given but it is claimed that the minimum energy detectable is 2×10^{-6} erg. The active area of the receiver was estimated as 1 mm^2. Here then is a bolometer of equivalent noise input not markedly inferior to the best thermo-couples but with a time constant hundred times better. The superiority of this bolometer to all other existing thermo-detectors has been confirmed independently by tests in another laboratory (Bell *et al.* 1946). It should be added that the whole unit, which weighs twelve and a half pounds, is such that when freshly primed with liquid hydrogen and liquid nitrogen it will operate for 4 hours or more without any adjustment. Furthermore, the same workers have developed a portable hydrogen liquifier (weighing 50 lb.) of the Ahlberg type, which will produce enough liquid hydrogen in two hours to operate the cryostat. This makes the super-conducting bolometer a practicable piece of apparatus for many laboratories and it should find a wide field of application.

Another novel type of bolometer is the dielectric a.c. bolometer first proposed by Moon and Steinhardt (1938). By following change of dielectric constant (instead of resistance) with temperature it would appear that much higher responsivities can be obtained than with metal or semi-conductor bolometers. However, the performance of the actual dielectric bolometers constructed by Moon and Steinhardt fell far below the theoretical figure. Recently Ewles (1947) has tried to develop this type of bolometer using nitrobenzene as the dielectric, but the performance is still markedly inferior to that of the thermistor bolometer.

The ultimate theoretical sensitivity of a bolometer is controlled by temperature fluctuations exactly as described above for the thermocouple. However, for the thermocouple we pointed out that in practice the equivalent noise input was nearly always determined by the Johnson noise in the circuit. The bolometer differs in two principal respects. The first is that in addition to Johnson noise we have current noise (Brillouin 1934) and possibly contact noise (Christiansen and Pearson 1936); the second is that bolometers are now becoming so rapid in response that the dependence of the expression for temperature fluctuation noise on the frequency of the signal has to be considered. This problem has been considered by Daunt (1945 b) who concentrates his attention in this paper on the relative magnitude of Johnson noise and temperature fluctuation noise. He shows that at high frequencies (above about 50 c.p.s.) the ratio of the Johnson noise to the temperature fluctuation noise increases very rapidly. According to Daunt the minimum detectable power at very low frequencies is not limited by Johnson noise, yet in practice it is our opinion that for very low frequencies the best vacuum thermocouple is superior to the best metal or thermistor bolometer. Since we

know that the thermocouple does not reach the limit set by temperature fluctuations by a considerable factor, there must be some other source of noise (current or contact) which is limiting the practical performance of all bolometers used at room temperature. We would emphasize that a comprehensive treatment of all the factors limiting the equivalent noise input of bolometers has yet to be published. The influence of backing the bolometer element with some materials (such as rock salt) to act as a "heat sink" has been considered by Shive (1947) and by Billings, Hyde and Barr (1947 b). From both these investigations it would appear that at high signal frequencies (>100 c.p.s.) backing does not contribute to and may even adversely affect the sensitivity of the bolometer. In certain applications, however, backing may be useful in reducing microphonic noise.

(viii) *Pneumatic detector*

An entirely different type of thermo-detector which was brought to a high degree of perfection during the war is the pneumatic heat detector of Golay and Zahl (Zahl and Golay 1946, Golay 1947 a, b). This was developed from an earlier detector of Hayes (1936) in which radiation falling on carbonized vegetable fluff enclosed in a small cell caused adsorbed gas to be emitted. The resulting expansion in the cell, one side of which consisted of a thin metal diaphragm, changed the electrical capacity of the cell. One of the main drawbacks in this type of detector from the service point of view was the clumsiness of the associated electrical amplifying gear when a large battery of detectors had to be assembled. Accordingly Golay and Zahl, while retaining the principle of the Hayes cell, modified it in two main respects. They replaced the carbonized fluff with a thin blackened collodion membrane, having satisfied themselves that the essential action of the fluff was to absorb infra-red radiation and raise the temperature of the surrounding gas rather than to exude gas. In the second place, they arranged an optical method of detecting and amplifying the distortion of the cell caused by the expansion of the gas. Full details are given in the most recent paper by Golay (1947 b) and the cell is now being manufactured by the Eppley Laboratory Inc. at Newport, Rhode Island, U.S.A. A typical cell is described by Golay as having "a time constant of 3 milliseconds and an equivalent noise input of about $1{\cdot}4 \times 10^{-9}$ watt". Unfortunately the band-width is not quoted. It is pointed out that the speed could be increased to give a time constant of 0·6 millisecond at the expense of a drop in sensitivity by replacing the air in the cell by helium. It would appear therefore that the Golay cell is comparable with the B.T.L. thermistor bolometer in performance, i.e. speed of response and equivalent noise input, but insufficient details have as yet been published to enable us to make an exact evaluation of their relative merits. Golay (1947 a) has discussed some of the theoretical aspects of his pneumatic heat detector in an interesting paper which stresses the useful analogies which can be drawn between thermal and electrical circuits. He concludes that the theoretical efficiency of the detector has been fairly closely approached.

(ix) *Radiometric*

There have been no new important developments in the technique of detection by radiometers during the past ten years, unless we regard the pneumatic detector described in (viii) above, as a form of radiometer.

(x) *Evaporographic and convective*

A method of detecting infra-red radiation which has come to be known as "evaporography" was introduced by Czerny (Czerny and Mollet 1937 a, b) shortly before the war. The radiation is allowed to fall on a film of celluloid covered with a thin film of paraffin, 0·01 μ in thickness. The celluloid is backed with aluminium to increase the efficiency of absorption. The whole is mounted in a vacuum of 0·01 mm. of mercury (i.e. vapour pressure of the paraffin at room temperature). The action of the radiation is to warm the paraffin and cause it to evaporate, so producing discontinuities in the thickness of the layer. These variations can be seen or photographed under suitable illumination by an interference technique. This method has been used successfully for recording spectra up to 10 μ. Some further details have been given by Czerny and Roder (1938). The ultimate sensitivity of this method is not known, but it would appear to have possibilities where picture forming rather than the highest sensitivity is important.

An interesting variation of this technique has been made recently by Heintz (1946). Instead of using the evaporation produced in a thin layer of paraffin, Heintz uses the convection currents by localized heating in a thin film of liquid. The principle is illustrated in figure 10, in which A represents a point image produced by infra-red rays in the middle of the film of liquid L. Suspended in the liquid are very fine particles of aluminium powder (about 0·01 mm. in diameter) which follow the convection currents indicated. After a short time (varying from a few seconds to a few minutes) an accumulation of the aluminium particles is produced at D. If, now, the whole cell is illuminated from the side, D will show up as a dark spot on a light background, the latter being produced by reflection from the aluminium particles which gradually settle out evenly from the suspension. One liquid used by Heintz with considerable success was amyl alcohol; another was a solution of picein in petrol. When the cell is placed in the exit focal plane of an infra-red spectrometer with a continuous source, the spectrum of the liquid in the cell is automatically recorded in a few seconds. This has been accomplished out to 7 μ but the resolving power is at present very poor. Heintz claims that the sensitivity is equal to that of a thermocouple, but the thermocouple he employed is very poor by present day standards, the minimum detectable power being quoted as $2{\cdot}3 \times 10^{-7}$ watt with a galvanometer of 7-seconds period.

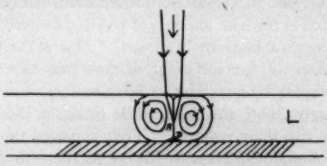


Figure 10. Production of infra-red picture by convective currents (Heintz).

Heintz has also explored a second method of observing the convection. This depends on the fact that a minute depression is produced in the surface of the liquid above A and that this depression can be measured by optical means. The prospect of obtaining satisfactory pictures in this way seems small, and unless a method of this type will give a good picture, it seems unlikely that it can compete with those based on a thermocouple or bolometer.

§3. NEW MATERIALS FOR INFRA-RED WORK

(i) *Dispersive materials*

Before 1939 the customary dispersive materials for infra-red work were glass, quartz, fluorite, rock salt, sylvine, potassium bromide and potassium iodide, according to the region under investigation (Sutherland 1935). We have arranged these in order of increasing transmission into the infra-red, quartz cutting off at about 3·6 μ and a potassium iodide prism allowing observations to be made up to about 32 μ. To obtain optimum results each of these materials should be used only from the cut-off of the preceding one to its own limit of transmission. However, this would necessitate many changes of prism to cover the whole range from the visible to 32 μ, and it used to be extremely difficult to obtain natural crystals of some of these materials sufficiently perfect to provide large prisms. Two recent developments have simplified this problem considerably; the first is the commercial production of large synthetic crystals of rock salt and potassium bromide, and the second is the introduction of synthetic lithium fluoride into the above series.

Synthetic crystals of rock salt and potassium bromide were originally made by slow crystallization from the melt by Pohl (1933) and Stockbarger (1936, 1937). The former used what is known as the "pot method" in which a seed crystal is introduced near the top of a melt in a cylindrical pot. The temperature of the pot is controlled very carefully and as the large crystal grows from the seed with approximately the same diameter as the pot, it is slowly withdrawn. In Stockbarger's method the whole melt is crystallized *in situ* by extremely slow cooling through the melting point. The latter method has been developed on a commercial scale by the Harshaw Chemical Company (Kremers 1940) and synthetic crystals of rock salt and potassium bromide can now be obtained quite easily, from which it is possible to cut a 60° prism of height 10–11 cm. and with a base of at least 15 cm. The perfection of these crystals is such that they are nearly always superior in two respects to a natural crystal of the same size. The average natural crystal has imperfections which cause beam distortion in the spectrometer with a consequent loss of definition and resolving power. Secondly, the natural crystals frequently contain traces of impurity (e.g. CaCl) which make them much more hygroscopic than the synthetic crystals.

The new material, lithium fluoride, is also a synthetic product by the Harshaw Company. Its value lies in the fact that it transmits up to about 6 μ, and in the range from 2·8 to 6 μ it is superior to fluorite in dispersive power (Wright 1944). Furthermore, it is more useful than quartz between 2·8 and 3·1 μ as the latter has strong absorption near 2·9 μ. Since large quartz and fluorite* prisms are extremely rare, lithium fluoride is a very valuable addition to the series of dispersive materials. It can be obtained as a 60° prism of any size up to 11 cm. high × 15 cm. base. Thus a very satisfactory prism spectrometer can now be made using lithium fluoride from the visible to 6 μ, rock salt from 6 μ to 15 μ and potassium bromide from 15 μ to 28 μ.

(ii) *Window materials*

Rock salt has been the customary window material up to 17 μ and for longer wavelengths, potassium bromide or iodide. However, all these materials are soft

* Since this was written, we have learnt that synthetic fluorite is also now available from the Harshaw Company.

and very susceptible to moisture. Recent work by the firm of Steinheil in Munich has shown that ground and polished rock salt plates can be made considerably more resistant to attack by moisture by a simple heating process (Elliott 1946). The poor resistance normally shown by a polished rock salt surface is ascribed to minute scratches and surface cracks, and it is claimed that these can be removed without affecting the optical quality by gradually heating the specimen to 500° C. in an oven.

During the war, sheet silver chloride has been developed on a commercial basis by the Harshaw Chemical Company of Cleveland, Ohio, U.S.A. Silver chloride is particularly valuable in infra-red work because of its resistance to moisture, its high transmission (80 % out to 20 μ) and its property, if melted at 457·5° C., of wetting glass and metals. According to Kremers (1947), thin sheets can easily be made by pressing very large single crystals into flat ingots. One of the principal defects of silver chloride is that it is decomposed (and so rendered opaque to infra-red radiation by the silver produced) when exposed to ultra-violet light such as that present in sunlight. However, it can be protected against such solarization by coating with silver sulphide and rolling the latter into the layer, although it should be added that this reduces the transmission very appreciably for wavelengths below 4 μ. Coating with selenium (Pfund 1933) or gilsonite is also useful in this respect. Silver chloride corrodes base metals, but noble metals, such as silver, gold and platinum, are not affected. Nickel, chrome plate and stainless steel are also fairly resistant. This means that optical parts such as lenses and absorption cells can be made by impact pressing with chrome-plated dies.

Under the direction of Dr. Smakula of Zeiss, a series of synthetic crystals have been developed which are of value for infra-red work. Two of these (known as KRS 5 and KRS 6) are of particular importance. The former, which was manufactured in considerable quantities for military purposes, has the composition

Table 3. Refractive index of KRS 5 and of KRS 6

Wavelength (μ)	Refractive index		Wavelength (μ)	Refractive index	
	KRS 5	KRS 6		KRS 5	KRS 6
0·589	2·628	2·337	2·8	2·398	2·200
0·66	2·577	2·310	3·0	2·397	2·199
0·78	2·510	2·270	3·5	2·394	2·197
0·90	2·479	2·251	4·0	2·393	2·196
1·0	2·461	2·240	4·5	2·391	2·194
1·1	2·448	2·232	5·0	2·390	2·193
1·2	2·439	2·226	6·0	2·389	2·190
1·3	2·431	2·221	8·0	2·387	2·187
1·4	2·425	2·218	8·0	2·386	2·184
1·5	2·420	2·215	9·0	2·384	2·180
1·6	2·416	2·213	10·0	2·281	2·176
1·7	2·413	2·210	11·0	2·379	2·172
1·8	2·411	2·209	12·0	2·377	2·167
1·9	2·408	2·207	13·0	2·374	2·162
2·0	2·406	2·206	14·0	2·371	2·156
2·2	2·404	2·204	15·0	2·369	2·150
2·4	2·401	2·203	16·0	2·366	2·144
2·6	2·399	2·201	17·0	2·364	2·128

42% thallium bromide and 58% thallium iodide, while the latter has 60% thallium chloride and 40% thallium bromide. They are both harder and far more moisture resistant than the alkali halides and have good transmission properties in the infra-red. A 2-mm. thick plate of KRS 5 has a uniform transmission of about 70% between 1 μ and 40 μ, and a similar specimen of KRS 6 has rather better transmission to about 30 μ. This transmission can be increased to about 95% over a limited wavelength region by suitably "blooming" the surfaces. Each melts at about 412° c. Refractive index values for these materials are given in table 3.

(iii) *Filters*

A considerable amount of research was done during the war on suitable filters in the infra-red, especially for the region between the visible and 3 to 4 μ. The results of some of these investigations have been published (Blout and others 1946, Shenk and others 1946) and a wide range of useful new filters is now available in the short wave region. The British work is contained in Admiralty Reports (*A.R.L. Reports* E130, 1939–46). In the longer wave regions the only advance over the usual filter materials such as quartz, lithium fluoride, mica and fluorite is one due to White (1947) who has pointed out that echelette diffraction gratings can be used to give broad band filters. Their action consists in diffracting out the shorter wavelengths and specularly reflecting all wavelengths above a critical value which depends on the grating spacing.

§4. APPLICATIONS

(i) *General*

The developments in infra-red technique described above were conditioned almost entirely by wartime applications. These were of two principal types (*a*) military and (*b*) spectroscopic. The services required means of signalling and of detecting targets either by the thermal radiation which they emit or by the radiation they reflect when illuminated by an infra-red searchlight. This necessitated a picture such as that given by the image converter, but the resulting restriction to wavelengths below 1·2 μ meant that only targets at relatively high temperatures and short distances could be seen pictorially. In order to detect targets at temperatures only slightly higher than that of the background it was essential to use a detector which was sensitive to wavelengths between 2 and 13 μ in conjunction with some form of scanning. The demand was therefore for a detector combining very high speed with very high sensitivity over this region. The spectroscopic applications of infra-red technique to the analysis of petroleum, synthetic rubber, plastics and chemicals made an exactly similar demand. Here it was vital to be able to record spectra at high speed and yet under fairly high resolving power, particularly between 2 and 15μ.

Now that the detectors are becoming generally available, it is clear that a wide field of application exists, beyond that contemplated at the time of development. In this section we discuss some of the more interesting of these applications (including wartime) and indicate others still to be made. It will be convenient to separate the spectroscopic from the more general applications, and deal with them in the next section.

Image converters were used during the war in a number of roles. Some instruments (e.g. Type U receivers) were designed for navigational use. Equipment was also used for recognition purposes, for vehicle driving at night and, in conjunction with small infra-red searchlights, to replace gunsights for firing at targets invisible to the unaided eye. The photograph in figure 11 was taken in

Figure 11. A military vehicle illuminated with infra-red radiation as seen through image converter, Type BR/E (R.G. receiver).

complete darkness through one half of a British binocular driving equipment (known as Type BR/E receiver) incorporating an R.G. tube. All these applications involved the development of special filters to absorb the visible radiation from the source. These filters had to be capable of withstanding high temperatures (arising from their proximity to high wattage lamps) and exposure to all kinds of meteorological conditions without deterioration (cf. § 3 (iii)).

Image converters were employed by the Germans for the location of aircraft. In this case the radiation is emitted by the target itself. Aeroplane exhausts usually reach a temperature of 400–500° c., and because of the rapid rise of the black body emission curve as it passes from the visible region to the near infra-red, they can be detected with image converters at considerable distances when completely invisible to the unaided eye. The equipment was not very successful in this application, however, as the exhaust systems could be shielded, and the radiation does not penetrate through fog or cloud.

There are a number of fairly obvious peace-time uses for image converters. The behaviour of animals and birds during darkness, or under conditions of low illumination, can be investigated. Unexposed photographic film can be inspected and manipulated without fogging. Refractometers can be extended to a range of $1{\cdot}2\,\mu$. Astronomers have a rapid means of comparing the near infra-red emission of stars. It is very probable that the ability of the image converter to detect radiation from surfaces at temperatures too low for detection by the eye will be of value in physical and engineering laboratories. With cooled cathodes this temperature is as low as 200° c.

In the rapid examination of old manuscripts and pictures by near infra-red radiation, the image converter is likely to replace the photographic method in much of the preliminary sifting of material. Thus the striking results hitherto obtained by the slower photographic process (Clarke 1941) are likely to be greatly extended. The same applies to applications of infra-red photography in the biological field (Lecomte 1943).

Infra-red phosphors have so far proved to be less sensitive than image converters and, as they cover a very similar spectral range, at first sight would appear to have no advantages. However, when plenty of energy is available, so that high sensitivity is unimportant, the phosphors should be much more convenient as they do not require a source of high voltage but only a mercury vapour lamp or even an ordinary incandescent filament lamp to "sensitize" them. Moreover, a phosphor has been made which can be stimulated with an ordinary incandescent lamp and produce one microlambert for several hours before losing half its sensitivity. When stored in the dark it keeps its sensitivity for several days (Urbach and others 1946). This may have important applications in x-ray radiography (Berg and Kaiser 1947). Ellickson (1946) has shown that $2{\cdot}6\times10^{16}$ light quanta per c.c. may be stored in a strontium sulphide phosphor activated with cerium and samarium.

The only other methods of detection which yield infra-red pictures without the use of some scanning mechanism are the evaporographic and the convective. These are inferior to the image converter and to phosphors in that the image takes an appreciable time to build up (usually at least a second) and even longer to be removed. On the other hand, their spectral range is much more extensive (to about $10\,\mu$) and we are thus provided with a potential means of taking infra-red pictures using wavelengths covering a fair part of the region of fundamental absorption bands arising from internal molecular vibrations. This may be important in biological applications. In the picturing of objects of low surface temperature where high speed in following rapid changes is not vital, where extreme sensitivity is not necessary and where the complications introduced by scanning with a bolometer are undesirable, these methods may find applications.

Coming next to the photoconductive cells, these have the same general field of application as bolometers and thermocouples. They are greatly superior in speed and sensitivity but inferior in being restricted to short ranges of wavelengths in the near infra-red and in having a non-uniform response within that range. It should be added, however, that lack of response to longer wavelengths may be an advantage in certain applications, e.g. where insensitivity to low temperature fluctuations is desirable. Besides the obvious military uses for the detection of targets such as

ships' funnels and aircraft exhausts, there are a growing number of other non-spectroscopic applications. One of the earliest and most important is that made by Lee and Parker (1946) in pyrometry. Here for the first time the possibility was demonstrated of measuring and recording rapid variations of surface temperature down to 100° C. In this connection, the use of these cells to detect the localized heating (hot spots) produced by friction (Bowden and others 1936, 1947) is likely to be of considerable importance.

In astronomy, new fields will be opened with these detectors not only because of their high sensitivity but because the radiation from extremely distant stars is scattered so much before it reaches the earth that for the most distant the proportion of infra-red radiation to visible and ultra-violet is increasingly great. Work along these lines has been started at Yerkes Observatory by Kuiper and his colleagues (Kuiper, Wilson and Cashman 1947, Kuiper 1947) and improved spectra already obtained of certain stars and of the atmospheres of the planets.* When used in conjunction with some form of scanning mechanism, photoconductive cells could be made to yield pictures using any desired wavelength of radiation between 1 and 5 μ. To get reasonable definition, however, the size of the sensitive receiving area will have to be reduced considerably, and it is not yet certain to what extent this will reduce sensitivity.

There are at present no essentially new applications of the improved bolometers and thermocouples which are not spectroscopic. The pneumatic detector (which is of course equivalent to a fast bolometer) has been used in conjunction with scanning devices to produce an infra-red "picture" of an aeroplane in flight (Golay 1947 b).

(ii) *Spectroscopic*

The spectroscopic applications of the new detectors are controlled primarily by the range of their spectral response. Image converters and phosphors will replace the photographic plate in the short wavelength region in much spectroscopic work. Where the highest sensitivity is required and where speed in following changes in not important, the photographic plate will still be best for work out to 1·3 μ. For rapid visual examination of spectra, the image converter is probably the most suitable detector out to 1·2 μ. However phosphors may be more convenient provided intensities are sufficiently high; furthermore, their spectral range is greater, Paul (1946) having reported successful spectrograms out to 1·53 μ.

The new photoconductive cells (more particularly lead sulphide and lead telluride) open up a wide new field in spectroscopy. In the first place their high sensitivity offers the possibility of greatly increased resolving power compared with what was attainable before the war with the best thermocouple, and (apart from the super-conducting bolometer) the wartime advances in the thermocouple and bolometer performance have been in speed, rather than sensitivity. Some indication of these possibilities has been given by Sutherland, Blackwell and Fellgett (1946) who were able to resolve lines in the water vapour spectrum near 2 μ which are only 0·14 cm^{-1} apart. It should be possible to reduce this by a factor of about 2 when better diffraction gratings are available. As the pre-war

* We are indebted to Dr. Kuiper for kindly communicating many of his results to us before publication.

limit of resolving power in this region was about $0{\cdot}6\,cm^{-1}$, the ultimate improvement in the near infra-red should be nearly tenfold. This will be extremely important in future applications of infra-red spectroscopy to the determination of molecular dimensions. Before the war the number of molecules for which moments of inertia (and hence internuclear distances) could be determined from the spacings in rotational fine structure was limited to about 25, because of lack of resolving power. This number should be increased enormously and very high precision will be possible in the measurements of the constants of the lighter molecules. It is to be hoped that the range of photoconductive cells can be extended beyond the present limit of about $5\,\mu$ since this restricts the number of bands which are accessible to this increased resolving power. However, provided the molecule under examination has one hydrogen atom, then at least one fundamental frequency should be observable at a wavelength between $2{\cdot}7$ and $5\,\mu$. It is for this reason that lead telluride cells are bound to be of more value to the spectroscopist than lead sulphide cells. The latter fall off very steeply in sensitivity beyond $3\,\mu$ and relatively few fundamental vibrations occur below $3\,\mu$.

Quite apart from the advantages these photoconductive cells have over the older detectors in sensitivity, their much higher speed of response also opens up new fields to the spectroscopist. It should be possible to present a steady picture of a reasonable portion of the infra-red spectrum below $5\,\mu$ on a cathode-ray screen with quite high resolving power. Thus spectra which are changing rapidly in time could be recorded by cinephotography. This is certain to have important applications, e.g. in the field of chemical kinetics.

As indicated earlier, the principal improvement in the performance of thermocouples and bolometers (with the exception of the super-conducting bolometer) has been in the direction of speed rather than sensitivity. It is accordingly possible to record spectra much more rapidly than before, particularly when the highest resolving power is not imperative. Using the B.T.L. thermistor bolometer, Daly and Sutherland (1947) have been able to present pictures 2 to $4\,\mu$ in extent, of the infra-red spectrum between 1 and 16μ on a cathode-ray screen, the time of the scan being of the order of 10 seconds (figure 12). This has great advantages in the observation of spectra which are changing with time although, as mentioned above, the substitution of a photoconductive cell for the bolometer now being made will allow much higher speeds of scanning in the region between 1 and 5μ. Similarly the use of a super-conducting bolometer would permit a considerable, though smaller, increase in scanning speed over the whole region of the spectrum.

The use of the evaporographic and convective methods of detection to produce pictures of spectra has already been mentioned (§ 2(x)). These methods are in some respects much simpler than the scanning method, but until more is known of their ultimate sensitivity, it seems that they can only be applied where rather low resolving power is acceptable and relatively slow changes are occurring in the spectra under observation. Furthermore, the accurate measurement of intensities and contours by such methods presents quite formidable difficulties.

In addition to the developments made possible by the new detectors in the pictorial display and higher resolution of spectra, there have been several less spectacular but very important improvements in technique. Automatic recording of spectra has entirely replaced point by point manual plotting, even for work at

the highest resolving power (Bell and others 1947). The troubles arising from atmospheric absorption due to water vapour and carbon dioxide have been eliminated by the use of double-beam spectrometers (Hardy 1939, Sweeney 1944, Sutherland and Thompson 1945, Wright and Herscher 1947) or by evacuation of the spectrometer (Bell and others 1947). Before the war, many of the major difficulties in infra-red spectroscopy arose from the use of a very sensitive galvanometer system to detect the extremely small E.M.F. from the thermocouple. Instability arose from mechanical and thermal disturbances. These difficulties have been largely removed by the elimination of the galvanometer. A direct current valve amplifier suitable for direct coupling to a low resistance thermocouple has been developed (Liston and others 1946) and manufactured by General Motors Corporation, Detroit, U.S.A. A more satisfactory method is to chop the radiation falling on the detector and employ an A.C. valve amplifier, tuned to the chopping frequency. This eliminates "drift" due to slow thermal changes,

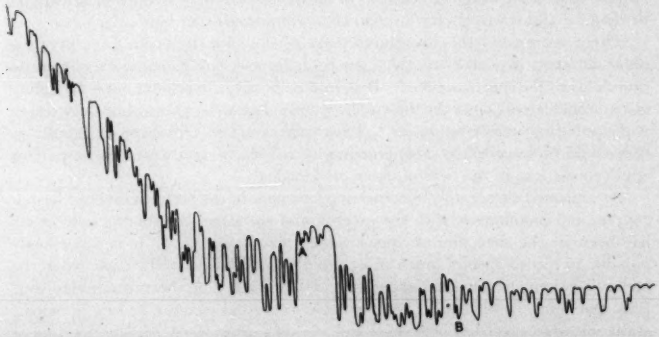


Figure 13. Fine structure in atmospheric water vapour band at 6μ as revealed by precision prism spectrometer (Blackwell and Sutherland). The separation of the doublet (A) is 2·2 cm^{-1} and in the triplet (B) the spacings are 2·8 cm^{-1} and 1·4 cm^{-1}.

which is still a drawback with a d.c. amplifier. For photoconductive cells and fast bolometers, where chopping frequencies between 20 and 1500 c.p.s. are possible, the design of such an amplifier presents no special problem, but for the more sensitive thermocouples, a chopping frequency of 5 to 10 c.p.s. is the upper limit. Special amplifiers have been devised for this purpose (e.g. Roess and Dacus 1945 b).

Improvement in the resolving power of prism spectrometers arising from the use of the new synthetic crystals has been demonstrated by several workers (Wright 1944, Barnes and others 1945), while more recently the relative merits of lithium fluoride and calcium fluoride have been carefully compared (Gore and others 1947). The excellence of the resolving power attainable with a well designed prism spectrometer (1 to 2 cm^{-1} separations in the range 6 to 15 μ) is well illustrated by the work of Nielsen, Crawford and Smith (1947) and by figure 13, taken from a recent spectrometer record in Cambridge. Many of the

above spectroscopic developments have been discussed in greater detail by one of us elsewhere (Sutherland 1947) together with a few of the applications to chemical problems.

The two major problems to which infra-red spectroscopy was applied during the war were the analysis of petroleum and the structure of penicillin. It would be beyond the scope of this article even to attempt to summarize the results. A review of the main results in the petroleum field should be available shortly (Sutherland 1948) and it is hoped that full accounts of the British and American work will not be long delayed. In the case of penicillin, a joint account of the British and American work will be available soon in the official book on penicillin now in course of publication. More comprehensive reviews of the applications of infra-red spectroscopy to chemistry have been given by Barnes, Gore, Liddel and Williams (1944) and by Thompson (1946).

ACKNOWLEDGMENTS

We are indebted to the Superintendent of the Admiralty Research Laboratory for permission to include in the section on photocells many unpublished results obtained in the laboratory, and to Mr. E. F. Daly for helpful discussion and criticism of the sections on thermocouples and bolometers.

REFERENCES

Admiralty Research Laboratory Reports, 1939–46. (*a*) ARL/E. 300 ; (*b*) ARL/E. 130; (*c*) ARL/E. 320 ; 1947. ARL/R. 8, 9, 10/E. 320.

Aiken, C. B., Carter, W. H., and Phillips, F. S., 1946. *Rev. Sci. Instrum.*, **17**, 377.

Allekotte, 1942. *German military report.*

Andrews, D. H., Milton, R. M. and de Sorbo, W., 1946. *J. Opt. Soc. Amer.*, **36**, 518.

Barnes, R. B., Gore, R. C., Liddel, U. and Williams, V.Z., 1944. *Infra-red Spectroscopy* (New York : Reinhold).

Becker, J. A., Green, C. B. and Pearson, G. L., 1946. *Elect. Engng.*, **65**, 711.

Becker, J. A. and Moore, H. R. 1946. *J. Opt. Soc. Amer.*, **36**, 354.

Bell, E. E., Buhl, R. F., Nielsen, A. H. and Nielsen, H. H., 1946. *J. Opt. Soc. Amer.*, **36**, 353.

Bell, E. E., Noble, R. H. and Nielsen, H. H., 1947. *Rev. Sci. Instrum.*, **18**, 48.

Berg, O. E. and Kaiser, H. F. 1947. *J. Appl. Phys.*, **18**, 343.

Billings, B. H., Barr, E. E. and Hyde, W. L. 1947 a. *Rev. Sci. Instrum.*, **18**, 429 ; 1947 b. *J. Opt. Soc. Amer.*, **37**, 123.

B.I.O.S., 1946. *Final Report* No. 530, Item 9.

Blackwell, D. E., Simpson, O. and Sutherland, G. B. B. M., 1947. *Nature, Lond.*, **160**, 793 ; 1948. *Nature, Lond.* (in press).

Blout, E. R., Amon, W. E., Shepherd, R. G., Thomas, A., West, C. D. and Land, E. H., 1946. *J. Opt. Soc. Amer.*, **36**, 460.

Bose, J. C., 1904. U.S.A. Patent No. 755840.

Bowden, F. P. and Ridler, K. E. W., 1936. *Proc. Roy. Soc.* A, **154**, 640.

Bowden, F. P. and Stone, M. A., 1947. *Proc. Roy. Soc.* A, **188**, 329.

Brattain, W. H. and Becker, J. A., 1946. *J. Opt. Soc. Amer.*, **36**, 354.

Brillouin, L., 1934. *Helv. Phys. Acta*, **7**, Suppl. 47.

Cartwright, C. H., 1934. *Z. Phys.*, **92**, 153.

Case, T. W., 1920. *Phys. Rev.*, **15**, 289.

Cashman, R. J., 1946. *J. Opt. Soc. Amer.*, **36**, 336.

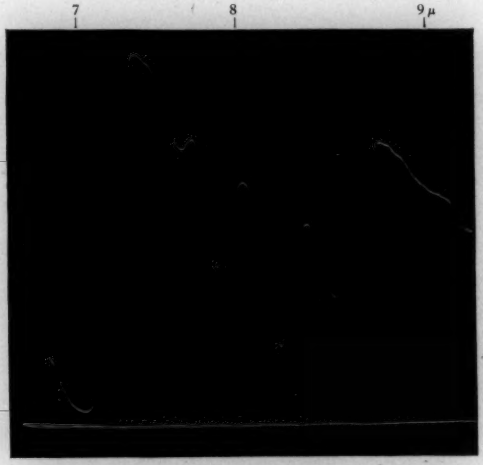
Christiansen, B. J. and Pearson, G. L., 1936. *Bell Syst. Tech. J.*, **15**, 197.

Clark, W., 1941. *Photography by Infra-red* (London : Chapman and Hall).

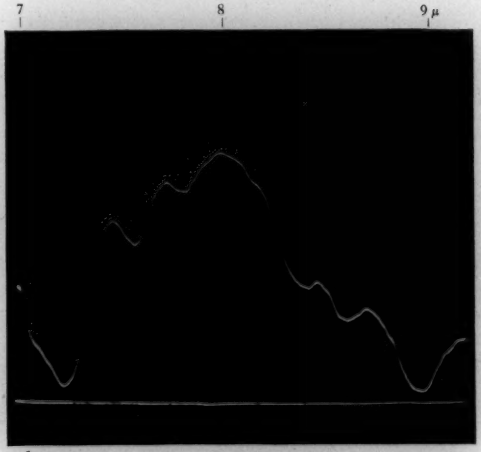
Czerny, M. and Mollet, P., 1937 a. *Z. Phys.*, **108**, 85; 1937 b. *Ibid.*, **108**, 582.

Czerny, M. and Roder, H., 1938. *Ergebn. exakt. Naturw.*, **17**, 70.

DALY, E. F. and SUTHERLAND, G. B. B. M., 1947. *Proc. Phys. Soc.*, **59**, 77.
DALY, E. F. and SUTHERLAND, G. B. B. M. Unpublished work done in 1940.
DAUNT, J. G., 1945 a. *Admiralty Report C.L. Misc.* 52; 1945 b. *Admiralty Report C.L. Misc.* 54.
ELECTRONIC ENGINEERING, October 1946.
ELLICKSON, R. T., 1946. *J. Opt. Soc. Amer.*, **36**, 501.
ELLIOTT, A., 1946. *Nature, Lond.*, **157**, 299; 1947. *Photo-cells for the Infra-red* in "*Electronics*" (London : Pilot Press).
EWLES, J., 1947. *J. Sci. Instrum.*, **24**, 57.
FIRESTONE, F. A., 1930. *Rev. Sci. Instrum.*, **1**, 360.
FISCHER, F., GUDDEN, B. and TREU, M., 1938. *Phys. Z.*, **39**, 127.
FONDA, G. R., 1946. *J. Opt. Soc. Amer.*, **36**, 501.
GARLICK, G. F. J. and WILKINS, M. H. F., 1945. *Proc. Roy. Soc.* A, **184**, 408.
GARLICK, G. F. J. and GIBSON, A. F., 1947. *Proc. Roy. Soc.* A, **188**, 485.
GOLAY, M. J. E., 1947 a. *Rev. Sci. Instrum.*, **18**, 347; 1947 b. *Ibid.*, **18**, 357.
GORE, R. C., MCDONALD, R. S., WILLIAMS, V. Z. and WHITE, J. U., 1947. *J. Opt. Soc. Amer.*, **37**, 23.
GRUETZMACHER, M., 1930. *Telegr.- u. Fernsprtech.*, **19**, 285.
HARDY, J. D., 1939. *Phys. Rev.*, **55**, 1112.
HARRIS, L., 1934. *Phys. Rev.*, **45**, 635 ; 1946. *J. Opt. Soc. Amer.*, **36**, 597.
HARRIS, L. and JOHNSON, E. A., 1933. *Rev. Sci. Instrum.*, **4**, 454 ; 1934. *Ibid.*, **5**, 153.
HARRIS, L., and SCHOLP, A. C., 1940. *J. Opt. Soc. Amer.*, **30**, 159.
HAYES, H. V., 1936. *Rev. Sci. Instrum.*, **7**, 202.
HEINTZ, E., 1946. *J. Phys. Radium VIII*, **7**, 293 and 336.
HEWLETT, C. W., 1947. *Gen. Elect. Rev.*, **50**, April, 22.
HOLST, G., DE BOER, J. H., TEVES, M. C. and VEENEMANS, C. F., 1934. *Physica*, **1**, 297.
HORN, F. H., BRUCKSCH, W. F., ZIEGLER, W. T. and ANDREWS, D. H., 1942. *Phys. Rev.*, **61**, 738.
HORNIG, D. F. and O'KEEFE, B. J., 1947. *Rev. Sci. Instrum.*, **18**, 474.
JOHNSON, J. B., 1928. *Phys. Rev.*, **32**, 97.
KUIPER, G. P., 1947. *Astrophys. J.*, **106**, 251.
KUIPER, G. P., WILSON, W. and CASHMAN, R. J., 1947. *Astrophys. J.*, **106**, 243.
KOLOMIEZ, B., 1938. *C.R. Acad. Sci., U.R.S.S.*, **19**, 5.
KREMERS, H. C., 1940. *Industr. Engng. Chem.*, **32**, 1478.
KREMERS, H. C., 1947. *J. Opt. Soc. Amer.*, **37**, 337.
LANGE, B., 1930. *Phys. Z.*, **31**, 964.
LECOMTE, J., 1943. *Le Spectre infrarouge et ses Applications dans les Sciences naturelles et biologiques* (Paris : Herman et Cie).
LEE, E. and PARKER, R. C., 1946. *Nature, Lond.*, **158**, 518.
LEWIS, W. B., 1947. *Proc. Phys. Soc.*, **59**, 34.
LISTON, M. D., QUINN, C. E., SARGEANT, W. E. and SCOTT, G. G., 1946. *Rev. Sci. Instrum.*, **17**, 194.
MACFARLANE, G. G., 1947. *Proc. Phys. Soc.*, **59**, 366.
MAJORANA, R. and TODESCO, G., 1928. *Atti Accad. d. Lincei VIII*, 9.
MILATZ, J. M. W. and VAN DER WELDEN, H. A., 1943. *Physica*, **10**, 369.
MILLER, N. F. and BARNETT, C. E., 1946. *J. Opt. Soc. Amer.*, **36**, 352.
MILTON, R. M., 1946. *Chemical Reviews*, **39**, 419.
MOON, P. and STEINHARDT, L. R., 1938. *J. Opt. Soc. Amer.*, **28**, 148.
MORTON, G. A. and FLORY, L. E., 1946. *Electronics*, Sept., p. 112.
MOSS, T. S., 1947. *Nature, Lond.*, **159**, 476.
MYERS, L. M., 1939. *Electron Optics* (London : Chapman and Hall Ltd.).
NIELSEN, J. R., CRAWFORD, F. W. and SMITH, D. C., 1947. *J. Opt. Soc. Amer.*, **37**, 296.
NIX, E. C. and TREPTOW, A. W., 1939. *J. Opt. Soc. Amer.*, **29**, 457.
OXLEY, C. L., 1946. *J. Opt. Soc. Amer.*, **36**, 356.
PAUL, F. W., 1946. *J. Opt. Soc. Amer.*, **36**, 175 and 352.
PFUND, A. H., 1933. *J. Opt. Soc. Amer.*, **23**, 375.
PFUND, A. H., 1937. *Rev. Sci. Instrum.*, **8**, 417.
PRATT, T. H., 1947. *J. Sci. Instrum.*, **24**, 312.



(*a*) 2.2.4 Trimethylpentane.



(*b*) 2.3.4 Trimethylpentane.

Figure 12. Infra-red spectra of hydrocarbons displayed on cathode-ray screen (Daly and Sutherland).

To face page 176.

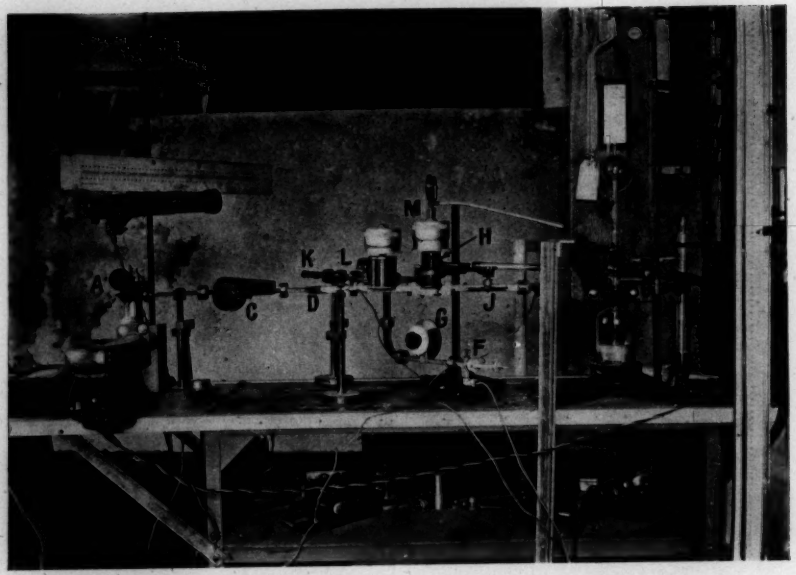


Figure 1. The apparatus.

A. Reflection klystron oscillator.
B. Air blower.
C, G, J, L. Variable attenuators.
D. Waveguide "twist".
E. Cavity wavemeter.
F, K. Crystal rectifiers.
H. Resonant cavity (experimental chamber).
M. Micrometer head.

ROESS, L. C. and DACUS, E. N., 1945 a. *Rev. Sci. Instrum.*, *R6*, 164 ; 1945 b. *Ibid.*, **16**, 172.
SCHOTTKY, W., 1926. *Phys. Rev.*, **28**, 74.
SHENK, J. H., HODGE, E. S., MORRIS, R. J., PICKETT, E. E. and BRODE, W. R., 1946. *J. Opt Soc. Amer.*, **36**, 569.
SHIVE, J. N., 1947. *J. Appl. Phys.*, **18**, 398.
SIMPSON, O., 1947. *Nature, Lond.*, **160**, 791.
SOMMER, A., 1946. *Photoelectric Cells* (London : Methuen and Co.).
SOSNOWSKI, L., SOOLE, B. W. and STARKIEWICZ, J., 1947. *Nature, Lond.*, **160**, 471.
SOSNOWSKI, L., STARKIEWICZ, J. and SIMPSON, O., 1947. *Nature, Lond.*, **159**, 818.
STARKIEWICZ, J., 1944. Br. Pat. 3571/44 and Br. Pat. 3572/44.
STARKIEWICZ, J., SOSNOWSKI, L. and SIMPSON, O., 1946. *Nature, Lond.*, **158**, 28.
STOCKBARGER, D. C., 1936. *Rev. Sci. Instrum.*, **7**, 133 ; 1937. *J. Opt. Soc. Amer.*, **27**, 416.
STRONG, J., 1939. *Procedures on Experimental Physics* (New York : Prentice Hall).
SUTHERLAND, G. B. B. M., 1935. *Infra-red and Raman Spectra* (London : Methuen and Co.) ; 1947. *Instruments and Measurements* (Stockholm, in the press) ; 1948. *Science of Petroleum*, Vol. X (Oxford, in the press).
SUTHERLAND, G. B. B. M., BLACKWELL, D. E. and FELLGETT, P. B., 1946. *Nature, Lond.*, **158**, 873.
SUTHERLAND, G. B. B. M. and THOMPSON, H. W., 1945. *Trans. Faraday Soc.*, **41**, 174.
SWEENEY, W. J., 1944. *Industr. Engng. Chem.*, **16**, 723.
THOMPSON, H. W., 1946. *Ann. Reports Chem. Soc.* for 1945, 1.
URBACH, F., PEARLMAN, D. and HAMMENDINGEN, H., 1946. *J. Opt. Soc. Amer.*, **36**, 377.
WRIGHT, N., 1944. *Rev. Sci. Instrum.*, **15**, 22.
WRIGHT, N., and HERSCHER, L. W., 1947. *J. Opt. Soc. Amer.*, **37**, 211.
ZAHL, H. A., and GOLAY, M. J. E., 1946. *Rev. Sci. Instrum.*, **17**, 511.

THE RADIO-FREQUENCY SPECTROSCOPY OF GASES

By B. BLEANEY,
Clarendon Laboratory, Oxford

§1. INTRODUCTION

THE extension of spectroscopy into new regions of the electromagnetic waves has usually been conditioned by the development of new sources and detectors of radiation. At first sight this would not appear to be the case at radio frequencies, where powerful sources and sensitive detectors have been available for many years. The subject matter of this Report consists, however, almost entirely of work carried out in the past three years, and published in 1946–7. This is due to the fact that the region of electromagnetic waves under investigation is that of centimetre wavelengths, for which adequate generators and detectors of oscillations have only been developed during the war. Their use for peacetime spectroscopic research has therefore only become possible very recently.

The importance of radio waves to the spectroscopist derives from the fact that the quantum of energy associated with the wave-packet is so small. For a wavelength of 1 cm. the quantum $h\nu$ is 2×10^{-16} ergs, and corresponds to a temperature of 1·4° K. It is thus possible to obtain from radio-frequency spectra information about the fine structure of the energy levels of atoms and molecules, which would otherwise require the use of the very highest resolving powers in measurements at optical or infra-red wavelengths. In atomic spectra the main interest lies in investigation of the hyperfine structure associated with the magnetic moment of the nucleus; in molecular spectra, besides the pure rotational spectra of the heavier molecules, there are fine structures associated with such effects as inversion, and with the electric quadrupole moment of the nucleus.

The smallness of the energy quantum is the main source of difficulty for the radio-frequency spectroscopist. The emission from a gas at a given temperature cannot exceed that from a black body at the same temperature, and lies therefore at the limit of detection of the most sensitive radio receivers. Moreover, if the gas is contained in a enclosure with conducting walls at the same temperature, the radiation therein will be black-body radiation, irrespective of the gas, and the emission spectrum of the latter will therefore be unobservable. In the case of the water vapour of the atmosphere, however, the emission has been measured by means of an ingenious radiometer due to Dicke (1946).

Apart from this one case, all observations on radio-frequency spectra have been made in absorption. The absorption coefficients are small, and no gas is opaque to centimetre waves to a degree comparable with absorptions in the optical region. The small energy quantum $h\nu$ plays a double rôle; not only is the actual energy absorbed in each transition small, but the numbers of molecules occupying the upper and lower energy levels between which a transition is

occurring are very nearly equal, and the absorption is nearly cancelled by the induced emission. Owing to the fact that $h\nu \ll kT$, the difference of populations is practically proportional to $h\nu/kT$, and the absorption coefficients vary with the square of the frequency of the transition. Thus with existing techniques it is only at centimetre wavelengths that the absorption due to gaseous spectra is easily detectable.

The problem of measuring the absorption can best be appreciated from a consideration of the actual magnitude of the absorption coefficients involved. In ammonia at atmospheric pressure, radiation of $1\frac{1}{4}$-cm. wavelength is absorbed to the extent of about 50 % in a metre of path, and it is therefore possible to measure the absorption accurately using moderate path lengths. Ammonia is, however, an exceptional case, and the absorption due to a line in a pure rotation spectrum is generally smaller by a factor of at least 100. Path lengths up to thirty feet have been used by some experimenters to obtain a measurable amount of absorption, but the more accurate methods have generally made use of the properties of the resonant cavity. The particular advantage of the latter may be regarded in two ways. The cavity may be regarded as a tuned circuit in which the energy stored is large, and the dissipation of energy in resistive losses small; the extra dissipation caused by the admission of an absorbing gas is therefore easy to detect. Alternatively, the wave in a highly resonant cavity may be regarded as traversing it many times before its final decay, and a large effective multiplication of path length is thereby obtained.

The differences in technique of the radio-frequency and the optical spectroscopist are, however, far larger than would be suggested by the difference in magnitude of the absorption coefficients. The sources of radiation, the detectors and the means of collimating the radiation are quite different in nature, corresponding to the fact that the wavelengths are now macroscopic rather than microscopic. The sources of oscillation are vacuum tube generators, and the radiation is, for practical purposes, monochromatic; no instrument of the nature of a spectroscope for isolating different wavelengths is therefore required. Collimation of the radiation by means of mirrors or lenses is impracticable, or at any rate highly inconvenient, because of diffraction effects. Instead a truly parallel beam of radiation, of cross-section less than one wavelength square, may be produced by means of waveguides or co-axial cables.

In view of the novelty of these techniques, a considerable portion of this report has been devoted to their description: in §2 a brief survey is given of general techniques of measurement with centimetre waves, and in §3 a description of the various methods hitherto developed for observing the spectra of gases at these wavelengths.

Apart from the intensity, the most important factor common to all the spectra is the breadth of the lines, which at ordinary pressures is due entirely to collision broadening. The mean time between collisions in a gas at atmospheric pressure is of the order of 10^{-10} seconds, which is comparable with the period of the radiation. It follows that the wave-trains emitted and absorbed will contain only a few waves, and the corresponding spectral lines are very broad. In these circumstances the collision broadening theory developed by Lorentz (1916) for the narrow lines of optical spectra needs revision, and Van Vleck and

Weisskopf (1945) have developed a rigorous theory, based on the concept of the classical harmonic oscillator, which contains the Debye (1929) theory of non-resonant absorption as a special case. Collision broadening plays such an important rôle in centimetre-wave spectra that the theory is outlined in fair detail (§4) before the spectra themselves are discussed.

The first interest in absorption spectra at centimetre wavelengths arose from the use of these waves for radiolocation, and it is therefore natural that a considerable amount of work, both theoretical and experimental, should have been devoted to the gases of the atmosphere. Of these, water vapour and oxygen produce absorption, the former due to its electric dipole and the latter due to its magnetic dipole. Of other gases, ammonia has been studied extensively, and the fine structure of its inversion spectrum has been observed by several workers. Resolution of this structure is only obtained by using very low pressures (down to 0·01 mm. Hg) when the line widths due to collision broadening are a few megacycles/second or less. These spectra are described in §5.

The predominance of the effects of collision broadening at centimetre wavelengths has enabled an extensive study of the phenomenon to be made; the value of such a study is considerably enhanced by the fact that measurements of absorption coefficients can be made with considerably greater precision than is usually achieved in the infra-red waveband. The experimental work on collision broadening (chiefly in the ammonia spectrum) is described in the last section (§6) of this Report.

§2. GENERAL EXPERIMENTAL TECHNIQUE

Except for the pioneer work of Cleeton and Williams in 1934, all experimental work on the radio-frequency spectroscopy of gases so far reported at centimetre wavelengths has been carried out using waveguide techniques. Since these techniques are considerably different both from those of the radio engineer dealing with wavelengths large compared with the dimensions of the apparatus and from those of the optical spectroscopist dealing with microscopic wavelengths, it seems desirable to give a general review of waveguide methods before considering in more detail their application to spectroscopy.

The transmission of electromagnetic waves inside hollow conductors is more complex than in free space or along the two-wire transmission line. If the dimensions of the waveguide are large compared with the wavelength of the radiation, many modes of propagation are possible, each with its own peculiar configuration of electric and magnetic fields. The phase velocity of the waves is different for each mode, depending on the ratio of the wavelength λ of the radiation in free space to a critical wavelength λ_c determined by the dimensions of the conductor and the field configuration for this mode. For a waveguide filled with a medium whose dielectric constant and permeability are unity, the ratio of the phase velocity v_p to the velocity in free space $\mathbf{c}$ is

$$v_p/\mathbf{c} = \{1-(\lambda/\lambda_c)^2\}^{-\frac{1}{2}}. \qquad \ldots\ldots(1\,a)$$

Thus v_p is greater than $\mathbf{c}$, and since v_p depends on λ, the propagation of the waves is subject to dispersion. The group velocity v_g is given by the formula

$$v_g/\mathbf{c} = \{1-(\lambda/\lambda_c)^2\}^{\frac{1}{2}}. \qquad \ldots\ldots(1\,b)$$

The attenuation of the wave depends on the ratio of the energy dissipated per

unit length to the energy transmitted; if the group velocity is smaller than that in free space, less energy is transmitted for a given energy density per unit volume, and hence for a given dissipation per unit length. Thus the attenuation of the wave caused by the presence of an absorbing medium is greater in a waveguide than in the unbounded medium, and depends, moreover, on the mode of propagation. It would therefore be necessary to know the proportion of energy transmitted in the different modes in order to deduce the true absorption coefficient of the medium. For this reason, and on account of the practical difficulties which would arise from the presence of several modes, a waveguide is used with only one mode of propagation excited. This is generally accomplished by choosing a size and shape of guide for which the critical wavelength of one mode only is greater than the wavelength of the radiation. Then if, for instance, a discontinuity in the guide causes the excitation of other modes, they cannot be propagated down the guide and are subject to a very rapid exponential decay. The most convenient shape of guide is rectangular, with the narrow dimension less than half a wavelength and the wide dimension between one-half and one wavelength. This allows the transmission of the H_{11}, or TE_{11}, mode. In this mode the electric field has only one component, parallel to the narrow side of the guide, while the magnetic field has components parallel to the wide side and in the direction of propagation. The waveguide shown in figure 1, is of this type, the inner dimensions being $1{\cdot}07 \times 0{\cdot}43$ cm., suitable for use at wavelengths of about 1·1 cm. to 1·8 cm.

Owing to the finite resistance of the conducting walls, the wave inside the guide is attenuated at a rate which depends on the material and dimensions of the wall and on the wavelength. At a wavelength of $1\frac{1}{4}$ cm., in a silver or copper waveguide of the dimensions just given, the theoretical rate of attenuation is about 0·4 decibels per metre; in practice it may be a little greater if the walls are not sufficiently smooth and clean. The variation with wavelength is proportional to $\lambda^{-\frac{1}{2}}$, if the guide is scaled in proportion.

The general technique of using waveguides for making electromagnetic measurements is fairly standardized, and may be illustrated by reference to a typical waveguide "test-bench" for wavelengths of about $1\frac{1}{4}$ cm. Figure 1* shows the apparatus used by Bleaney and Penrose (1948) for measurements on the ammonia spectrum between 1·1 cm. and 1·6 cm. wavelength. A reflex klystron oscillator A, cooled by an air blower B, feeds power into a rectangular waveguide excited in the H_{11} mode. The power level in the guide can be adjusted by a variable attenuator C, which introduces attenuation up to 20 db. with negligible reflection. At least 10 db. of attenuation is desirable to "decouple" the oscillator from the rest of the apparatus; otherwise, changes in the impedance of the latter may affect the frequency and power output from the oscillator in an unknown manner. For convenience, the plane of polarization of the radiation is rotated through 90° by the "twist" D; the latter illustrates the general principle that no reflection of power is caused by distortion of the waveguide, if the distortion is gradual, so that no sharp discontinuities occur in distances of less than a wavelength. The wavelength of the radiation is determined by a sharply resonant cavity wavemeter E, resonance being detected by a silicon-tungsten

* Plate facing page 177.

crystal rectifier F. The power level incident on the latter is adjusted by another variable attenuator G. The wavemeter absorbs a small amount of power from the guide when tuned to resonance but has no effect when detuned. The experimental chamber in these experiments was the resonant cavity H, excited by a small amount of power fed through small holes in the side of the guide. The main guide terminates in an attenuator J, followed by a crystal detector (not visible), which is useful as a monitor for the oscillator output, and for adjusting the oscillator to obtain the most steady conditions. Resonance in the experimental cavity is detected by a third crystal detector K, preceded by an attenuator L. The attenuators are always set to the maximum consistent with a reasonable power for detection by the crystals; this prevents large standing-wave patterns being set up in the guide, which might cause undesirable couplings between the different components.

This test bench comprises elements which are common to all waveguide measurement work. For a complete description reference should be made to wireless and electrical journals (e.g. "Proceedings of the Radiolocation Convention", March–May 1946, *J. Instn. Elect. Engrs.*, **93**, III A, 1946), and the following brief notes on the components must suffice for the present article.

Oscillators.—Although magnetrons have been used by some experimenters, the most useful generator is the reflex klystron. It gives a continuous wave output, ranging from about 20 milliwatts at wavelengths of 1 cm. to a few hundred milliwatts or more at 10 cm. wavelength. The tuning range of these tubes is about 5 %, although special types with a greater range are available at the longer wavelengths. Tubes are not available to cover the complete band of centimetre wavelengths, but the gaps could be filled by modifications of existing tubes.

Small power continuous wave magnetrons are available at the longer wavelengths, and may be used instead of klystrons, though they offer no special advantages. At certain wavelengths high power magnetrons are available, giving powers of the order of kilowatts in pulses of about one microsecond in duration, with a pulse repetition rate of 500 or 1000 per second. Practically all of these magnetrons are fixed in frequency, but, due to the short pulse, their output is spread over several megacycles per second in frequency; the frequency spread may be greater than that appropriate to the pulse length if the tube is not run under the optimum conditions of electrode voltages and load. Because of this spread pulsed tubes are unsuitable for observation of very narrow lines or use with sharply tuned resonators.

The range of wavelengths over which measurements can be made may be extended by the use of frequency multiplication. The harmonic content of the output from a reflex klystron is very small, but if the output is applied to a device such as a crystal rectifier whose characteristic is non-linear, an appreciable amount of harmonic power may be generated. By this means absorption measurements have been made by Beringer (1946) at a wavelength of about 5 mm. The method may also be used to fill in the gaps at longer wavelengths where generators are not available.

Detectors.—The most sensitive detector, and that generally used, is the silicon-tungsten crystal valve rectifier. This is available in the form of capsules

where a sensitive contact has been found in manufacture, and stabilized so that it is not disturbed by normal handling. The crystal rectifier may be used as a simple detector, giving a rectified current proportional to the incident radio-frequency (r.f.) power, or as a frequency converter giving an intermediate-frequency voltage proportional to the radio-frequency voltage. It is much more sensitive in the latter method, a signal power of the order of 10^{-14} watts being sufficient to give an output equal to the noise generated internally in the receiver if the band-width of the latter is one or two megacycles per second. This great sensitivity is often not required, and then the crystal may be used as a simple detector, giving a few microamperes of rectified current per microwatt of r.f. power. The rectified current may be measured by a galvanometer, which is less subject to fluctuations than an amplifier; it is not always accurately proportional to the r.f. power, and the crystals need to be checked at frequent intervals. It is preferable to use a low resistance galvanometer, as the crystal characteristic is generally more accurately "square-law" when it is working into a low-impedance load.

The crystal law can only be checked if some means of producing a known variation in the r.f. power is available, and some other detector with a law which is known or which may justifiably be assumed is generally required for this purpose. If sufficient power is available, the alternative detector may be used directly. Bleaney and Penrose (1947 *b*) have used a vacuum bolometer containing a fine steel wire the resistance of which is altered by the heating caused by the r.f. power. To avoid the danger of "burn-out" of the crystal by the high peak power from the magnetron, Becker and Autler (1946) used a series of thermo-couples, alternate junctions being coated with a high-loss mixture of iron powder in a polystyrene base. This absorbs the radiation and these junctions are the "hot" junctions; this procedure is necessary because the radiation is also incident on the "cold" junctions.

Measurement of wavelength and frequency.—Although frequency standards are available from the harmonics of quartz crystals, the measurement of wavelength is often more convenient. By the use of sharply resonant cavities, wavelengths may be measured with an accuracy limited by the machining tolerance in the making of the cavity and the micrometer head. The design of such cavities has been considered by Bleaney, Loubser and Penrose (1947), who describe a wavemeter for wavelengths of 1·1 cm. to 1·6 cm. with an accuracy of one or two parts in 10^4. With this instrument the wavelength in free space can be calculated from the dimensions of the cavity and the wavelength in the cavity. Other experimenters have used cavities calibrated at small frequency intervals against harmonics of a standard frequency, a continuous frequency coverage being obtained by interpolation.

Although some increase in the accuracy of the direct measurement of wavelength may be achieved, it seems unlikely that it will approach that of the direct measurement of frequency by means of the harmonics of quartz crystals; in recent measurements Strandberg, Kyhl, Wentink and Hillger (1947) and Good and Coles (1947) have attained accuracies of the order of 1 part in 10^6. Since the velocity of light is known only with an accuracy of about 1 part in 10^5, it becomes necessary to observe a distinction between wave-number (the reciprocal of the

wavelength in cm.), and frequency: the former should be reserved for observations where wavelength is measured directly, and the latter for observations where frequency is measured directly (including, of course, the case where resonant cavities are used to interpolate small frequency intervals between harmonics of a standard frequency). The most convenient unit, and that generally used to express microwave frequencies, is the megacycle/second (Mc/s); the writer would like to suggest, however, that the less clumsy term "megahertz" (MHertz) be used, as in the remainder of this report.

§3. SPECTROSCOPIC METHODS

3.1. *Emission Methods—The Dicke Radiometer*

The problem of measuring the thermal radiation at radio frequencies can be appreciated from a discussion of the performance of radio receivers. For quantitative consideration of the performance the concept of an "ideal receiver" has been found useful. The ideal receiver has been defined as a receiver which introduces no internal fluctuations or "noise"; its output consists of the amplified components of the fluctuating electromagnetic fields incident on the aerial, the latter being assumed to be in thermal equilibrium with an enclosure at room temperature T. The thermal radiation power intercepted by the aerial is then $kT\Delta\nu$, where $\Delta\nu$ is the band-width of the receiver. A practical receiver does introduce additional fluctuations, of course, due to shot noise in the valves and Johnson noise in the resistors, and its output contains the amplified components of these fluctuations, which thus obscure the "aerial noise". The consequent reduction in the performance is measured by the "noise factor", defined as the ratio of the total noise output to the amplified component of the aerial noise. Receivers for centimetre wavelengths have noise factors of the order of 10; that is, the total thermal radiation accepted by the aerial contributes only about 10 % of the noise output from the receiver. The difficulties of measuring small changes in the density of the thermal radiation are obvious.

The problem of increasing the output of the receiver due to the thermal radiation relative to that due to the internal fluctuations has been solved by Dicke (1946). The principle of the method is to modulate the amplitude of the incoming radiation at a low audio frequency (30 c/s.) by means of a rotating "shutter", and then to limit the band-width of the audio-frequency amplifying section of the receiver to a fraction of a cycle centred on 30 c/s. The output meter then records the 30 c/s. signal due to the modulation of the thermal radiation, together with those components of the internal fluctuations which lie within the pass-band of the audio-frequency amplifier, and final detector. The advantage of this technique is as follows: since the fluctuation power is proportional to the band-width, both for the internal fluctuations and for the thermal radiation, a simple limitation of the band-width affects both equally. The action of the shutter is to introduce a coherent component at 30 c/s. into the thermal radiation fluctuations, while the internal fluctuations remain incoherent. The enhancement of the output due to the former relative to that due to the latter is proportional to the square root of the ratio of the input band-width (16 MHertz) to the output band-width (*c.* 0·2 c/s.). Since the thermal radiation power is proportional to its absolute temperature, it is convenient to specify

the useful sensitivity of radiometer in degrees K. The useful sensitivity is defined as "the antenna temperature change required to equal the R.M.S. noise fluctuations in the output meter", and is measured by replacing the antenna by a "dummy load" of variable temperature; the sensitivity attained by Dicke is 0·4° K.

The nature of the "shutter" in this radiometer is important. It consists of an absorbing disc, which, on insertion into the waveguide, cuts off the receiver from the aerial. Since the disc is at room temperature, its action is to replace the external thermal radiation by a source of noise whose intensity equals that of thermal radiation of room temperature. Thus, strictly, the radiometer only detects radiation whose equivalent temperature is different from that of the disc.

3.2. *Absorption Methods*

3.21. *The waveguide transmission method.*—The most straightforward method of measuring absorption is to allow the radiation to flow down a waveguide which can either be evacuated or filled with gas. The absorption is then determined from the decrease in the intensity of the radiation reaching a detector placed at the far end of the waveguide. If the relation between the detector reading and the power incident upon it is accurately known, the most serious error arises generally from small reflections in the waveguide, which reduce the power reaching the detector. Owing to the comparatively small attenuation in the guide, multiple interference effects occur, which may alter on the admission of gas. Although the susceptibility of the gas is small, the path-lengths in the guide may contain many hundreds of wavelengths and an appreciable change of phase between the various reflections may be caused. Thus the detector reading may change even with a non-absorbing gas; it is therefore necessary to eliminate all reflections as far as possible, in order to obtain accurate measurements of absorption.

The most serious reflections arise at the vacuum windows closing the waveguide cell; these are therefore made of the thinnest possible mica, and several experimenters (Beringer 1946, Bleaney and Penrose 1947 b) have used two windows spaced a quarter-wave apart, the reflections from which should cancel if they are identical. The interference effect may also be reduced by introducing a considerable attenuation within the waveguide cell (Walter and Hershberger 1946).

Owing to the dispersive properties of a waveguide, the absorption of radiation transmitted through it is greater than that through an unbounded medium by the factor (equation (1 *b*))

$$c/v_g = (1-\lambda^2/\lambda_c^2)^{-\frac{1}{2}}.$$

This factor can be calculated from the wavelength of the radiation and the dimensions of the guide. The admission of a gas with a finite susceptibility causes an increase in λ_c, and hence a slight increase in the absorption due to ohmic losses in the guide; this effect is, however, generally negligible (see Hershberger 1946).

An interesting application of the waveguide transmission method is the display of very narrow spectral lines (obtained by working at gas pressures of the order of 0·1 mm. Hg) on a cathode-ray tube (Good 1946, Townes 1946). The frequency of the oscillator is varied at an audio-frequency rate, and the

rectified signal from the detector is amplified and made to produce a vertical deflection of the spot on the screen of a cathode-ray tube, the time-base of which is synchronized with the modulation of the oscillator frequency. The range over which the oscillator frequency can be varied is about 20 MHertz, and lines of widths up to about 10^{-4} cm^{-1} can be displayed. The accuracy of the method is such that absorption coefficients can be measured with an error of 2×10^{-5} per cm. (Townes 1946).

The method of displaying the spectral lines on a cathode-ray tube has the considerable merit of great rapidity of detection of absorption lines, and is therefore very useful in the preliminary analysis of a spectrum. Its sensitivity has recently been improved by Hughes and Wilson (1947) by the use of the radio-frequency Stark effect field, which modulates the absorption by the gas so that a radio receiver can be used for detection purposes. An alternating voltage at a frequency of 80 kHertz is applied to a central insulated electrode in the waveguide absorption cell; the electric field causes a splitting of the spectral line due to Stark effect, thus varying the absorption due to the gas. The centimetre-wave power reaching the detector is therefore modulated at 80 kHertz in the vicinity of an absorption line. The sensitivity is such that the line $(J, K) = (1, 1)$ of the N^{15}-ammonia inversion spectrum can be observed when present in the natural proportion of 0·3 % in ordinary ammonia.

Further improvements in technique for detecting weak lines have been reported by Gordy and Kessler (1947 b), and Watts and Williams (1947 b, c).

3.22. *Resonant cavity methods.*—The advantage of using a resonant cavity for the measurement of small absorptions has been mentioned in the introductory section. The method has been applied in two different ways—one using a cavity resonant in a single mode, the other a cavity resonant simultaneously in many modes. The approach to these methods is very different and it is convenient to consider them separately.

The single-mode cavity.—The cavity resonant in a single mode may be considered as a length of waveguide closed at each end by a reflecting wall. Since the ohmic losses are relatively small, the radiation traverses the cavity many times before its final decay and a large effective multiplication of path length is obtained. This gain is only achieved, however, if the multiply reflected waves are all in the correct phase relation to give reinforcement. The cavity is therefore a sharply tuned device, and its length must be adjusted to an exact number of half-wavelengths (in the guide) of the incident radiation.

This requirement of tuning makes the use of circular guides more convenient than rectangular, and Bleaney and Penrose have used a cylindrical cavity excited in the H_{01} (or TE_{01}) mode, tuned by a plunger driven by a micrometer head. Since the dimensions of the cavity are large enough to permit the propagation of several modes, selective methods of exciting the cavity are used (Bleaney, Loubser and Penrose 1947). The determination of absorption is made by measuring the circuit magnification factor "Q" (defined as 2π times the ratio of the energy stored to the energy dissipated per cycle) of the cavity when evacuated (Q_0) and when filled with the absorbing gas (Q_1). The absorption coefficient of the gas is then given by the expression

$$\alpha = (2\pi/\lambda)(1/Q_1 - 1/Q_0).$$

The "Q" of the cavity is determined from the resonance curve of the cavity by measuring the movement of the plunger required to detune the cavity to the half-power points. Owing to small fluctuations in the oscillator frequency, this measurement is accurate to within about 3 %, and the measurement of absorption coefficients by observing the change in "Q" directly will only be accurate when the absorption is comparable with the ohmic losses in the cavity, i.e. when Q_1 is materially different from Q_0. This limitation is overcome by measuring instead the change in the energy density in the resonator when the gas is admitted; this is done by observing the power reaching a loosely coupled detector. During measurement the cavity is kept on tune, and fluctuations in the oscillator frequency cause only second order fluctuations in the detector reading. The energy density is proportional to Q^2, and a change of 1 % can be detected. At wavelengths near 1 cm. the value of Q_0 is about 10 000; thus it is possible to detect absorptions of the order of 2×10^{-6} per cm. (1 decibel per km.).

The multiple-mode cavity.—Becker and Autler (1946) have used a copper cube, 8 ft. along the edge, for the measurement of the very small attenuation due to water vapour in the air at wavelengths between 0·7 cm. and 1·7 cm. Such a cavity is similar to the "hohlraum" used in the theory of thermal radiation, except that energy is being dissipated by ohmic losses on the walls etc. The rate of loss of energy is, of course, proportional to the total energy stored in the cavity, and conversely, if energy is supplied at a constant rate, the energy density in the cavity will be inversely proportional to the total losses. Thus the losses due to different absorbers may be compared by measuring the energy density by a detector with output proportional to the square of the average electric field in the cavity.

The theory of the instrument has been examined by Lamb (1946). An essential condition is that a large number of normal modes of the cavity must be excited to approximately the same degree. Experimentally this is ensured in two ways: (*a*) the cavity is excited by power from a magnetron in pulses of $\frac{1}{4}$ microsecond duration, (*b*) the input power is directed at the blades of a copper fan rotated at 2 to 3 r.p.s., and a second fan mounted near one of the side walls. These fans act as "mode-mixers" for the radiation, which is spread over a band of several megahertz corresponding to the Fourier analysis of the pulses; altogether some 66 000 modes were appreciably excited.

The mean energy density in the cavity, averaged both over the volume and over a time long compared with the interval between pulses, was determined by means of 360 pairs of chromal-constantan thermocouples distributed in strings across the cavity. The calibration of the instrument for the absolute measurement of absorption was obtained by the introduction of a known loss of energy through an aperture in the walls. Lamb has shown that if the dimensions of this aperture are small compared with those of the box, but still large enough to avoid diffraction effects, then its effect may be expressed as an additional damping $(1/Q_A)$ equal to

$$1/Q_A = \lambda A/8\pi V,$$

where λ is the wavelength of the radiation, A the area of the aperture and V the volume of the cavity. That the additional damping was in fact accurately proportional to the area of the aperture was confirmed experimentally, using two different openings. The Q of the box was found to be approximately 8×10^4

and absorptions of the order of 10^{-6} per cm. ($\frac{1}{2}$ decibel per km.) were measured with a probable error of 5 %.

§4. THEORY OF COLLISION BROADENING AND ABSORPTION

At all ordinary pressures the width of spectral lines at centimetre wavelengths is due to the effect of collisions between the molecules, which cause an interruption of the absorption of radiation, and hence, by limiting the wave-train, give rise to a broadening of the line. Lorentz (1916) showed that if the chance of a time t elapsing between collisions has a simple exponential probability distribution of the form

$$f(t) = (1/\tau)e^{-t/\tau},$$

where τ = mean time between collisions, the shape of the spectral line should be given by a structure factor

$$\Delta\nu/\{(\Delta\nu)^2 + (\nu_0 - \nu)^2\},$$

where
ν = frequency of applied radiation,
ν_0 = natural frequency of oscillator,
$\Delta\nu$ = "line breadth constant" $= 1/2\pi\tau$.

The line breadth constant is thus related to the mean time τ between collisions, and is equal to half the width of the line at half intensity.

Lorentz was concerned with the optical region, where the width of a line is very small compared with its frequency ($\Delta\nu \ll \nu_0$). This is far from being the case at centimetre wavelengths, and, as a preliminary to his work on absorption in the atmosphere, Van Vleck reviewed the theory of Lorentz. The final form of the revised theory is given in a paper by Van Vleck and Weisskopf (1945), where the theory of Lorentz for resonant absorption and that of Debye (1929) for non-resonant absorption are welded into a harmonious whole.

The theory of Van Vleck and Weisskopf is developed primarily for "strong collisions" (that is, collisions after which the values of the orientation, phase etc., of the molecule are independent of the values before collision), but is extended to "intermediate" and "weak" collisions for the Debye absorption. The major difference in the theory from that of Lorentz, is that the transient amplitudes of the oscillator after collision are assumed to be not perfectly random, but distributed in accordance with a Boltzmann law. Orientations or phases for which the molecule has a lower energy with respect to the field of the incident radiation are thus favoured, and the distribution corresponds to the maintenance of thermal equilibrium, whereas for the random distribution of Lorentz this was not the case.

Van Vleck and Weisskopf consider first the case of Debye absorption with non-resonant, rotating dipoles. If a collision takes place at an instant t_0 when the applied field is $E\cos\omega t_0$ and the dipole μ makes an angle ψ with the applied field, then the relevant Boltzmann factor by which the mean polarization after a collision is determined is $\exp(\mu E\cos\psi\cos\omega t_0/kT)$. Averaging over an exponential distribution of the mean times after the last collision, the polarization per unit volume is found to be

$$P = \frac{NE\mu^2}{3kT} \times \text{real part of } \left(\frac{e^{j\omega t}}{1 + j\omega\tau}\right). \qquad \ldots\ldots (2)$$

Hence the absorption coefficient α is

$$\alpha = \frac{\omega}{c}\frac{4\pi N\mu^2}{3kT}\frac{\omega\tau}{1+\omega^2\tau^2}; \qquad \ldots\ldots(3)$$

this is the same as the usual Debye expression with a relaxation time τ, which is here equal to the mean time between collisions.

The Lorentz theory of collision broadening is developed for a classical harmonic oscillator of charge e, mass m and natural frequency ω_0 for which the equation of motion is

$$m\left(\frac{d^2x}{dt^2}+\omega_0^2 x\right) = eE\cos\omega t,$$

of which the solution is

$$x = \text{real part of}\left\{\frac{eEe^{j\omega t}}{m(\omega_0^2-\omega^2)} + C_1 e^{j\omega_0 t} + C_2 e^{-j\omega_0 t}\right\}.$$

The transient amplitudes C_1, C_2 are determined by the values of x and dx/dt immediately after a collision. The assumption made by Lorentz was that, on the average, both x and dx/dt are zero after a collision. Instead, Van Vleck and Weisskopf assume that the values of x and dx/dt should be distributed according to the Boltzmann distribution law for a Hamiltonian function

$$H(t) = \tfrac{1}{2}m(dx/dt)^2 + \tfrac{1}{2}m(\omega_0 x)^2 - eEx\cos\omega t.$$

When the transient amplitudes are determined in this way, and the polarization is averaged over the times of last collision, one obtains for the absorption coefficient α and the refractive index n the expressions

$$\alpha = \frac{2\pi Ne^2}{mc}\left(\frac{\omega}{\omega_0}\right)\left\{\frac{1/\tau}{(\omega-\omega_0)^2+1/\tau^2}+\frac{1/\tau}{(\omega+\omega_0)^2+1/\tau^2}\right\}, \qquad \ldots\ldots(4\,a)$$

$$n^2-1 = \frac{4\pi Ne^2}{m(\omega_0^2-\omega^2)}\left[1-\frac{1+(\omega/\omega_0)}{2\tau^2[(\omega-\omega_0)^2+1/\tau^2]}-\frac{1-(\omega/\omega_0)}{2\tau^2[(\omega+\omega_0)^2+1/\tau^2]}\right]. \qquad \ldots\ldots(4\,b)$$

These results can be adapted to the quantum-mechanical system by replacing e^2/m by $8\pi^2\nu_{ij}|\mu_{ij}|^2/3h$, and ω_0 by $2\pi\nu_{ij}$, where $\nu_{ij}=(W_i-W_j)/h$ is the characteristic frequency associated with transition between the states i and j. One obtains then

$$\alpha = \frac{8\pi^3\nu}{3hc}|\mu_{ij}|^2 f(\nu_{ij},\nu)\left\{\frac{Ne^{-W_j/kT}}{\Sigma e^{-W_j/kT}}-\frac{Ne^{-W_i/kT}}{\Sigma e^{-W_j/kT}}\right\}, \qquad \ldots\ldots(5)$$

where

$$f(\nu_{ij},\nu) = \frac{1}{\pi}\frac{\nu}{\nu_{ij}}\left\{\frac{\Delta\nu}{\Delta\nu^2+(\nu_{ij}-\nu)^2}+\frac{\Delta\nu}{\Delta\nu^2+(\nu_{ij}+\nu)^2}\right\} \qquad \ldots\ldots(6\,a)$$

is called the "shape factor" of the line. The two terms $Ne^{-W_j/kT}/\Sigma e^{-W_j/kT}$ and $Ne^{-W_i/kT}/\Sigma e^{-W_j/kT}$ give the number of molecules in the lower and upper levels between which transitions are taking place, and the negative sign preceding the second term is due to the fact that the molecules in the upper level give induced emission which tends to cancel the absorption due to those in the lower level. At radio frequencies, $h\nu \ll kT$, and

$$(e^{-W_j/kT}-e^{-W_i/kT}) \simeq e^{-W_j/kT}(h\nu_{ij}/kT).$$

Hence

$$\alpha = \frac{8\pi^3\nu\nu_{ij}}{3kTc}|\mu_{ij}|^2 f(\nu_{ij},\nu)\frac{Ne^{-W_j/kT}}{\Sigma e^{-W_j/kT}}. \qquad \ldots\ldots(5\,a)$$

Bleaney and Penrose (1947 b) have suggested the use of a slightly modified nomenclature; a "structure factor" $F(\nu_{ij}, \nu)$ is defined as

$$F(\nu_{ij}, \nu) = \frac{1}{\pi}\left\{\frac{\Delta\nu}{\Delta\nu^2+(\nu_{ij}-\nu)^2} + \frac{\Delta\nu}{\Delta\nu^2+(\nu_{ij}+\nu)^2}\right\}, \quad \ldots\ldots(6\,b)$$

whence

$$\alpha = \frac{8\pi^3\nu^2}{3\boldsymbol{ckT}}|\mu_{ij}|^2 F(\nu_{ij}, \nu)\frac{Ne^{-W_i/\boldsymbol{kT}}}{\Sigma e^{-W_i/\boldsymbol{kT}}}. \quad \ldots\ldots(5\,b)$$

The advantage of this structure factor is that

$$\int_0^\infty F(\nu_{ij}, \nu)d\nu = 1,$$

from which it can be seen that the integrated intensity of the line

$$\int \frac{\alpha}{\nu^2}d\nu = \frac{8\pi^3|\mu_{ij}|^2}{3\boldsymbol{ckT}}\frac{Ne^{-W_i/\boldsymbol{kT}}}{\Sigma e^{-W_i/\boldsymbol{kT}}}. \quad \ldots\ldots(7)$$

Although the theory is developed for resonant absorption, non-resonant absorption should be included as a special case. On putting $\nu_{ij}=0$, and $\Sigma|\mu_{ij}|^2=\mu^2$, and remembering that $e^{-W_i/\boldsymbol{kT}}/\Sigma e^{-W_i/\boldsymbol{kT}}=\frac{1}{2}$ in this case, $(5\,a, b)$ reduces to

$$\alpha = \frac{8\pi^2\nu^2\mu^2 N}{3\boldsymbol{ckT}}\frac{\Delta\nu}{\Delta\nu^2+\nu^2} \quad \ldots\ldots(8)$$

which is the same as the Debye expression (3).

For narrow lines ($\Delta\nu \ll \nu_{ij}$) the second term in the structure factor, eqs. $(6\,a, b)$ can be neglected, and the modifications introduced into the theory of Lorentz are unimportant. For broad lines the extra term cannot be neglected, and its effect is considerable at frequencies far from resonance. In addition, the theory of Van Vleck and Weisskopf is very satisfactory by virtue of its success in unifying the theories of Lorentz and Debye. Fröhlich (1946) has arrived at the same results by an independent method, in which the effect of collisions is represented by an exponential return to the thermal equilibrium value of the distribution of the harmonic oscillators in displacement and momentum from the perturbed value caused by the absorption of radiation.

The absorption coefficient α in these equations is the attenuation in nepers per cm. of path. An alternative unit, often more convenient for the small absorptions encountered in the centimetre wave region, is the decibel/km. Values of absorption expressed in this latter unit are greater by a factor of $10^6 \log_{10}e$, than values expressed as attenuation per centimetre.

§5. THE SPECTRA

Owing to the great use of centimetre waves for radiolocation the first interest in absorption spectra was concerned with the gases of the atmosphere. In their passage through the atmosphere radio waves suffer attenuation due to scattering by small particles and to absorption. The latter was investigated theoretically by Van Vleck (1947 a, b). Absorption arises from both water vapour and oxygen—in the case of water vapour from a weak rotational line at 1·35 cm. wavelength, together with the "tails" of rotational lines further into the infra-red, in the case of oxygen from a resonant absorption at 0·5 cm. wavelength, together with a small non-resonant absorption, both associated with the magnetic dipole moment of the oxygen molecule. It is convenient to consider the two molecules separately.

5.1. *Water vapour*

The water molecule is an asymmetric top molecule, with an electric dipole moment of $1{\cdot}87 \times 10^{-18}$ e.s.u. Its rotational spectrum has been extensively investigated (for a review, see Dennison 1940). Most of the lines lie well into the infra-red, having wavelengths less than 0·1 cm.; there is, however, a weak line due to transition between two rather high rotational levels, called in the usual notation, 5_{-1} and 6_{-5}. According to the work of Randall, Dennison, Ginsberg and Weber (1937), slightly revised by Dennison (1940), these lie at 446·39 cm^{-1} and 447·17 cm^{-1} respectively, giving a line at 0·78 cm^{-1}; owing to the experimental difficulties of the infra-red measurements, the line was, however, only determined within the limits 1·0 cm^{-1} and 0·66 cm^{-1}, and its width was uncertain. The Boltzmann factor is 0·11, and the matrix element of the transition probability has been calculated by Van Vleck as 0·17 times the square of the dipole moment. If the line breadth constant is taken as 0·1 cm^{-1}, the attenuation in an atmosphere composed of 1 % of water vapour is 0·12 db/km. at the centre of the line. To this must be added a contribution of about 0·005 db/km. from the tails of other lines, the exact magnitude of which is again uncertain owing to lack of knowledge of the line widths.

The smallness of the absorption makes the task of experimental observation a difficult one. The principal measurements have been made by the Dicke radiometer (Dicke, Beringer, Kyhl and Vane 1946), and by the multiple-mode cavity (Becker and Autler 1946). In the former method, the thermal radiation from the atmosphere was measured at three wavelengths, 1·00, 1·25 and 1·50 cm., with the radiometer directed towards the zenith, and at angles of 48·2°, 60° and 66·5° to the vertical. Some cumulus clouds were found to emit radiation, and consequently the radiometer was never pointed at clouds during absorption measurements, nor, obviously, at the sun; cosmic noise was assumed to be negligible, and no evidence to the contrary was found. The humidity and temperature of the upper atmosphere were measured by balloon and aircraft soundings, and from these data the absorption due to the water-vapour line was deduced. The results are shown in table 1.

Table 1

Wavelength (cm.)	Absorption coefficient (db./km. per gm./cub. m.)		Total absorption in atmosphere due to oxygen (db.)
	Measured	Theoretical (at 1015 mb., 293° K.)	
1·00	0·011	0·0071	0·07
1·25	0·026	0·017	0·04
1·50	0·014	0·0092	0·03

The theoretical values assume the line centre to be at 1·33 cm. and the line breadth constant to be 0·12 cm^{-1}; the contribution from the tails of other lines is included. The absorption due to oxygen was found from extrapolation of the results to zero density of water vapour.

The eight-foot copper cube used by Becker and Autler as a multiple-mode cavity was maintained at 45° C. ± 1° C. so that a high vapour density could be

reached without condensation on the walls. A major difficulty arose from the fact that the cavity could not be evacuated; it was therefore necessary to use a circulation system of considerable capacity to ensure a uniform vapour density and temperature throughout the cavity, and to change the vapour density. The humidity was measured by means of wet and dry bulb thermometers in the vertical pipes of the circulation system. To obviate the necessity for long period stability in the source of power while the vapour density was being changed, the thermocouple E.M.F. was measured at constant density and constant power with the aperture in the walls of the cavity alternately closed and open. Thus the Q of the cavity is measured at a given vapour density; this procedure was repeated at densities up to 50 gm/cub.m., at eleven frequencies between $\bar{\nu}=0{\cdot}592$ cm^{-1} and 1·34 cm^{-1}. Contrary to what might have been expected, the absorption due to the water vapour did not increase linearly with the concentration; this can be seen from figure 2 where the specific absorption (db/km. per gm/cub.m.) is shown as a function of the wave-number, for two different concentrations, 10 gm/cub.m. and 50 gm/cub.m. At the centre of the absorption line the specific absorption diminishes slightly as the concentration increases, while off resonance it increases, particularly on the high-frequency side. This is consistent with an increase in line breadth with concentration, due to the cross-section for water–water collisions being considerably greater than the cross-section for water–air collisions. The ratio of the two cross-sections is found to be 4·7, the effect of the different velocities of water and air molecules being apparently neglected in this calculation (if it is included, the ratio of cross-sections is found to be about 10 % higher).

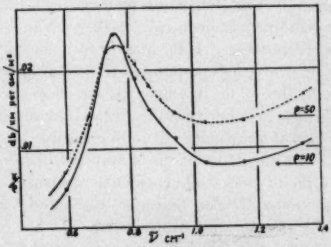


Figure 2. Water vapour absorption curve. The attenuation in db/kilometre per gram/cub.m. is plotted against the wave-number. The dotted curve is for a density of 50 gm/cub.m., the full curve for 10 gm/cub.m.

(*Reproduced, by kind permission, from 'Physical Review'*)

Extrapolation of the curves of absorption/concentration to zero density, and comparison with the theoretical work of Van Vleck, shows that the centre of the absorption line is at $\bar{\nu}=0{\cdot}744$ cm^{-1}, and its line breadth constant, $\Delta\bar{\nu}=0{\cdot}087$ cm^{-1}. The intensity is about 10 % higher than calculated, while the contribution from the tails of other lines is about four times greater than calculated. A similar comparison of the absorption curves for 50 gm/cub.m. of water vapour gives $\bar{\nu}=0{\cdot}742$ cm^{-1}, $\Delta\bar{\nu}=0{\cdot}107$ cm^{-1}; the intensity of the resonant absorption line is 13 % higher than calculated, and the tails some six times higher. The values of the line breadth constant are in good agreement with the value found by Adel (1947) of 0·11 cm^{-1} and 0·12 cm^{-1} for the water vapour lines at 15·99 μ and 18·64 μ respectively. The strength of the absorption found by Dicke and his co-workers is in agreement, within the experimental error, with the results of Becker and Autler.

Townes and Merritt (1946), using a waveguide transmission method with an oscillator swept in frequency across the line, have located the water vapour line at a pressure of about 0·1 mm. Hg of pure water vapour. They find its centre to be at 22237 ± 5 MHertz ($0{\cdot}7418 \pm 0{\cdot}0002$ cm^{-1}). The line breadth constant is 1·45 MHertz at a pressure of 0·103 mm. Hg, and the intensity at resonance $4 \pm 2 \times 10^{-6}$ nepers/cm. Townes and Merritt have also located one of the rotational lines of HDO predicted by Hainer, King and Cross (1946) to lie near one centimetre wavelength.

Saxton (1947) has measured the absorption in steam at wavelengths of 9 cm., 3·2 cm. and 1·6 cm. at several temperatures between 100° c. and 200° c. At the two longer wavelengths only an upper limit to the absorption could be determined, but at 1·6 cm. wavelength, the absorption coefficient at 100° c. and 760 mm. Hg was found to be 14·5 nepers/km. Though the difference in temperature makes comparison with other results difficult, one would expect that the absorption in steam due to the line at 0·742 cm^{-1} to be rather less than that found by Townes and Merritt ($4 \pm 2 \times 10^{-6}$ nepers/cm. at resonance). It is therefore probable that most of the absorption in steam must be ascribed to the tails of the infra-red lines.

5.2. *Oxygen*

The existence of absorption in the atmosphere due to oxygen was predicted in 1942 by Van Vleck (1947 a). The ground state of the oxygen molecule is $^3\Sigma$, with a magnetic dipole moment of 2 Bohr magnetons ($S=1$, $g=2$), which interacts with the rotation of the molecule to form a "rho-type triplet" (Schlapp 1937). The spin quantum number S compounds with the rotational quantum number K to form a resultant J, with values $K+1$, K, $K-1$. The components $K+1$, $K-1$ practically coincide, and differ from $J=K$ in energy by about 2 cm^{-1}. The allowed transitions are $\Delta J = 0, \pm 1$; the former gives rise to a non-resonant absorption, and the latter to a resonant absorption at wavelengths of about $\frac{1}{2}$ cm.

The strength of the non-resonant absorption is (in decibels per kilometre)

$$\gamma = (\log_{10} e) 10^6 \frac{8\pi^3 NS(S+1)\beta^2}{9\boldsymbol{k}Tc} \frac{\rho\nu^2}{1+\rho^2\nu^2}, \qquad \ldots\ldots(9)$$

where N = no. of molecules/cm^3.

β = Bohr magneton.

ρ = relaxation time.

For an atmosphere of the normal composition (fractional no. of molecules of oxygen = 0·21) at 293° K. and 760 mm.Hg this yields the value

$$\gamma = 0{\cdot}34 \Delta\bar{\nu}\,\bar{\nu}^2/(\Delta\bar{\nu}^2 + \bar{\nu}^2) \text{ db/km.}, \qquad \ldots\ldots(9\,a)$$

where $\Delta\bar{\nu} = 1/c\rho$. The actual value of $\Delta\bar{\nu}$ is uncertain, but is probably less than 0·1 cm^{-1}; the maximum absorption then occurs at a wavelength greater than 10 cm. and will not be more than 0·017 db/km.

At wavelengths shorter than this, the resonant absorption is much more important. The separations of the energy levels of the triplets are not known exactly, and depend on the rotational quantum number K. Van Vleck adopts the following values, based on the measurements of Dieke and Babcock (1927).

Table 2. Separation of the energy levels of the oxygen triplet

Rotational quantum no. K	Energy separation in cm^{-1}	
	$J=K$ to $J=K+1$	$J=K$ to $J=K-1$
1	1·87	3·96
3	1·95	2·08
5	1·98	2·00
7	2·02	1·97
9	2·04	1·94
11	2·06	1·92
13	2·08	1·90
15	2·10	1·88
17	2·12	1·86
19	2·14	1·84
21	2·16	1·82
23	2·18	1·81
25	2·20	1·80
>25	2·20+0·01 $(K-25)$	1·80−0·01 $(K-25)$

All the separations lie close to 2·0 cm^{-1} except that from $(J, K)=(1, 1)$ to $(0, 1)$; this transition has not been detected in the electronic bands, and the value 3·96 cm^{-1} has been calculated by Van Vleck from the theoretical formulae of Schlapp. Only odd values of the rotational quantum number appear in this table, since the nuclear spin of oxygen is zero, and consequently half the rotational levels are missing.

The computation of the absorption in the atmosphere involves lengthy numerical work, since the separation of the absorption lines is of the same order as their half width. The transition probabilities have the matrix elements

$$|\mu_{K,K+1}|^2=4\beta^2K(2K+3)/(K+1), \qquad \ldots\ldots(10\,a)$$

$$|\mu_{K,K-1}|^2=4\beta^2(K+1)(2K-1)/K. \qquad \ldots\ldots(10\,b)$$

The correct value of the line breadth constant is again uncertain, and Van Vleck uses values of $\Delta\bar{\nu}$ between 0·02 cm^{-1} and 0·1 cm^{-1}. The predicted atmospheric attenuation at a wavelength of 5·0 mm. is 12 to 14 db/km. for $\Delta\bar{\nu}=0{\cdot}02$ cm^{-1} and 9db/km. for $\Delta\bar{\nu}=0{\cdot}1$ cm^{-1}. The uncertainty in the value for narrower lines is due to the lack of knowledge of the exact location of individual lines; with the wider lines this is unimportant, since they would not be resolved.

The only experiments so far carried out on this oxygen spectrum have been those of Beringer (1946). Radiation of ½-cm. wavelength modulated at 1000 c/s. was generated from reflex klystrons oscillating at wavelengths between 0·96 cm. and 1·25 cm., using a silicon-tungsten crystal rectifier. The radiation traversed a rectangular waveguide 6·19 metres long (sealed with double mica windows at each end), which could be filled with the gas under investigation. The radiation was detected by specially prepared silicon-tungsten crystals, followed by a narrow-band 1000-c/s. amplifier.

Measurements were made using pure oxygen, air, and a 50–50 mixture of oxygen–nitrogen. The absorption per molecule of oxygen was found to depend only on the wavelength and the total pressure, indicating that the collision cross-section for oxygen–nitrogen collisions is substantially the same as for oxygen–

oxygen collisions. The measurements were not made at small enough pressures or sufficiently far from the centre of the absorption band to give an accurate value for the line breadth constant, but the strength of the absorption was in good agreement with that predicted by Van Vleck, if the line breadth constant ($\Delta\bar{\nu}$) is between 0·02 cm^{-1} and 0·05 cm^{-1}. This is appreciably smaller than the value of 0·087 cm^{-1} found for the water line in air, and is not much greater than that calculated from the kinetic theory cross-section of the oxygen molecule.

As pointed out by Beringer, the following points remain to be determined: (*a*) whether $\Delta\bar{\nu}$ is truly constant over the whole band of lines, and accurately linear with total pressure; (*b*) the form and location of the individual lines.

5.3. *Ammonia*

The form of the ammonia molecule is that of a rather flat pyramid with the nitrogen atom at the apex, and the three hydrogen atoms at the corners of an equilateral triangle as the base. The nitrogen has two equilibrium positions, one on either side of the plane of the hydrogen, and resonance between these two positions causes the doublet splitting well known in the infra-red spectrum of ammonia (for a review, see Dennison 1931, 1940). The most accurate measurement of this splitting in the ground vibrational state was that of Wright and Randall (1933), who showed that the pure rotation spectrum in the region of 100 μ consists of doublets with a separation of 1·33 cm^{-1}. The separation of the sub-levels of the rotation energy states is half of this, and since transitions between two sub-levels are allowed, ammonia should show an absorption near 0·67 cm^{-1}. The existence of this absorption was confirmed by Cleeton and Williams (1934), using split anode magnetrons as oscillators. The wavelength of the radiation was measured by means of a diffraction grating, and the absorption

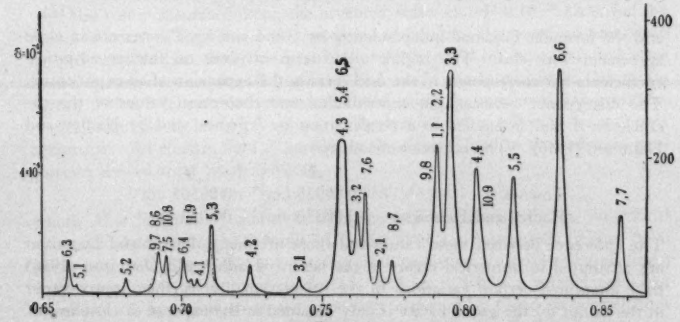


Figure 3.

(*Reproduced, by kind permission, from the* '*Proceedings of the Royal Society*'.)

was observed by placing ammonia (in a cloth bag at atmospheric pressure) in the path of a beam of radiation collimated by large mirrors. The peak of the absorption curve was found to lie at about 0·8 cm^{-1}, and the absorption was spread over a considerable range (0·5 cm^{-1}) owing to collision broadening.

New measurements of this ammonia spectrum have been carried out at low

pressure by Bleaney and Penrose (1946, 1947 a), Good (1946), and Townes (1946). The latter two workers have used waveguide transmission methods, the former a cavity resonant in a single (H_0) mode. The spectrum observed at a pressure of 1·2 mm. Hg is shown in figure 3. The extensive fine structure is attributed to the shift in the inversion frequency caused by centrifugal distortion of the molecule; thus each line in the inversion spectrum corresponds to a different rotational state. A splitting of this type had been observed previously by Sheng, Barker and Dennison (1941) in the 10 μ and 16 μ vibration-rotation bands of ammonia, and was shown by them to be in good agreement with their theoretical calculations. The inversion frequency is a delicate function of the height of the pyramid; rotation about the symmetry axis tends to decrease the height, as the hydrogen atoms tend to fly outwards. This makes it easier for the nitrogen atom to penetrate the potential barrier, and the frequency of inversion increases. On the other hand, rotation of the molecule "end-over-end" increases the height of the pyramid, and decreases the inversion frequency. Thus the splitting should follow a formula of the type

$$\bar{\nu} = \bar{\nu}_0 - A(J^2+J) + BK^2,$$

where J, K are the values of the two quantum numbers defining the rotational state of a symmetrical top molecule. The calculation of Sheng, Barker and Dennison predicted the values 0·0011 cm^{-1} and 0·0016 cm^{-1} for the coefficients A and B respectively in the ground vibrational state.

The splitting observed experimentally is considerably greater than this. The formula given by Bleaney and Penrose is

$$\bar{\nu} = 0{\cdot}7935 - 0{\cdot}00505(J^2+J) + 0{\cdot}00704K^2 + 0{\cdot}63\{-0{\cdot}0050(J^2+J) + 0{\cdot}0070K^2\}^2 \text{ cm}^{-1} \text{ (\textit{vacuo})}, \quad \ldots\ldots(11)$$

and the formulae obtained independently by Good and by Townes are in close agreement with this. The higher order term involves no further empirical coefficients but corresponds to the next term in the expansion of an exponential. The discrepancy between the experimental and theoretical values of the coefficients A and B has led to a recalculation by Townes, and by Hadley and Dennison (1946). The values obtained are

	A	B
Townes	0·00330 cm^{-1}	0·00505 cm^{-1}
Hadley and Dennison	0·00350 cm^{-1}	0·0053 cm^{-1}

The difference between these values and those of Sheng, Barker and Dennison are attributed to numerical errors in the latter. Hadley and Dennison (1946) have attempted certain variations in the calculation, but further improvement in the values for the ground state is only obtained at the expense of closeness of agreement with the experimental values for the first excited state. The discrepancies may be attributed partly to the approximate nature of the theory and partly to lack of knowledge of the details of the molecular forces in ammonia.

Precision measurements of the frequencies of the ammonia lines have recently been made by Strandberg, Kyhl, Wentink and Hillger (1947 a, b) and Good and Coles (1947). The accuracy claimed is about 1 part in 10^6, and the two sets of measurements are in agreement to this order. The former workers have

found it possible to fit the observed frequencies very closely by the formula (for $N^{14}H_3$)

$$\nu = \nu_0 \exp(\gamma/\nu_0) + \delta,$$

where ν = frequency in MHertz.

$\nu_0 = 23{,}785{,}75$ MHertz.

$\gamma = -151{\cdot}540(J^2+J-K^2)+59{\cdot}892K^2$ (MHertz).

$\delta = 0{\cdot}01353(J^2+J-K^2)^2+0{\cdot}00461K^2(J^2+J-K^2)-0{\cdot}00986K^4$ (MHertz).

The deviations of the observed frequencies from the calculated values are regular and less than 2 MHertz, except for the lines corresponding to $K=3$. Similar deviations for these lines have been found in the case of $N^{15}H_3$. This anomalous behaviour has been explained by Nielsen and Dennison (1947) as due to a splitting of the levels in the fourth order of approximation because of an interaction between the first-order Coriolis terms and first-order correction terms to the moments of inertia of the molecule. Symmetry considerations show that such a splitting can only occur when K is a multiple of three. The numerical values obtained by Nielsen and Dennison are in good agreement with the observed values for the $K=3$ lines; when $K=6$ or more the effect is very small.

The theoretical intensities of the lines in the inversion spectrum are given by equation (5 *b*) with the matrix elements of the dipole moment given by

$$|\mu_{ij}|^2 = |\mu_{J,K}|^2 = \mu^2K^2/(J^2+J).$$

The observed intensities are generally in agreement with the theoretica values. The most accurate measurements of the absorption coefficient are those of Bleaney and Penrose, who find that the integrated absorption between $\bar{\nu}=0{\cdot}67$ cm^{-1} and $0{\cdot}87$ cm^{-1} at a pressure of 4·5 mm. Hg agrees within 1 % with the value calculated using the accepted value of $1{\cdot}44\times10^{-18}$ E.S.U. for the dipole moment of ammonia. At a pressure of 0·5 mm. Hg seventeen lines were so well resolved that their line breadth constants could be measured, and were found to vary between 2×10^{-4} and 5×10^{-4} cm^{-1}; the intensities of these lines agreed within 5 % with the values calculated using the measured line breadths.

Coles and Good (1946) have observed Stark and Zeeman effects in the ammonia spectrum. An electric field E parallel to the electric field of the radiation produces a second order Stark splitting

$$\bar{\nu}\,(\text{cm}^{-1}) = 1{\cdot}5\,.\,10^{-4}\{MK/(J^2+J)\}^2\,.\,E^2,$$

where M = "magnetic" quantum number, which agrees with the theoretical value calculated from the formula of Polder (1942). A magnetic field of 6600 gauss perpendicular to the H.F. electric field produced a doublet (on all lines) with a separation of approximately $2{\cdot}0\,.\,10^{-4}$ cm^{-1}, while a parallel field had no effect. A hyperfine structure previously reported by Good on several lines is ascribed to the effect of the electric quadrupole moment of N^{14}; ammonia with the N^{15} isotope shows no such splitting, as is expected from the fact that the spin of N^{15} is $\frac{1}{2}$, while that of N^{14} is 1. The hyperfine structure has been investigated independently by Dailey, Kyhl, Strandberg, Van Vleck and Wilson (1946), who show that the observed splittings in $N^{14}H_3$ for the lines $(J, K)=(1, 1)$, (2, 2), (3, 3), (4, 4), and (3, 2) are in accord with their theoretical predictions, while no splitting is found with $N^{15}H_3$. The frequencies of the inversion lines

of the latter are, according to Coles and Good, given approximately by the formula

$$\bar{\nu}(\text{cm}^{-1}) = 0{\cdot}7575 - 0{\cdot}0047((J^2+J) + 0{\cdot}0068K^2.$$

Further observations of the hyperfine structure of $N^{14}H_3$ have been made by Watts and Williams (1947 a), Gordy and Kessler (1947 a), Williams (1947) and Jen (1947). The Stark splitting of the hyperfine structure has been investigated theoretically by Jauch (1947).

5.4. *Other molecules*

The inversion phenomenon is not peculiar to ammonia, but is possible with all non-planar molecules. In general, however, the potential hill dividing the two equilibrium configurations is so large that the inversion frequency is very low. Thus in phosphine this frequency has been calculated (Sutherland, Lee and Wu 1939) to be 4·5 MHertz, and in heavier molecules it will be considerably lower (for a discussion, see Herzberg, *Infra-red and Raman Spectra of Polyatomic Molecules* (Van Nostrand), p. 224). These gases, if polar, will show absorption at high frequencies, however, when the pressure is so great that the collision frequency is comparable with the frequency of the radiation. The shape of the absorption curve to be expected is similar to that for the Debye non-resonant absorption; when the inversion frequency is very small compared with the frequency of the applied radiation, one may write $\nu_{ij} = 0$ in the theory of Van Vleck and Weisskopf (§ 4, eq. (5 *a*)). The absorption is then given by eq. (8), except that the matrix elements correspond to the square of that component of the dipole moment μ which is stationary in space, summed over all rotational states, rather than to the square of μ itself. The presence of absorption due to this cause has been verified in a number of gases (six halogenated methanes and other organic gases) at wavelengths of 3·18 cm. and 1·24 cm. by Hershberger (1946) and Walter and Hershberger (1946).

Though most heavy polar molecules should have pure rotation lines in the centimetre wavelength region, these are generally so small in intensity for the simpler molecules as to lie at the limit of detection by present methods. This is due mainly to the very small fractions of the molecules occupying the low energy levels between which centimetre wavelength transitions can take place. This fraction is greater for linear than for more complicated molecules, and several rotational lines of linear molecules have been identified. Dakin, Coles and Good (1946) have located the second rotational line of carbon oxy-sulphide, OCS, at 24320 MHertz, from which the moment of inertia is calculated to be $1{\cdot}379.10^{-38}$ gm/cm². In an electric field of 2 to 3 E.S.U. a splitting due to a second order Stark effect is observed, from which the dipole moment is estimated as $0{\cdot}72.10^{-18}$ E.S.U. Recently, Dakin, Good and Coles (1947) have observed the corresponding line for OCS with the S^{34} isotope, and from this and more accurate measurements with the S^{32} isotope, the moment of inertia of OCS^{32} is found to be $(137{\cdot}9540 \pm 0{\cdot}0001)10^{-40}$ gm.cm². The bond distances are found to be (in the lowest vibrational state)

$$\text{C–O}: 1{\cdot}1612 \pm 0{\cdot}0058 \text{ A.}$$
$$\text{C–S}: 1{\cdot}56040 \pm {\cdot}0049 \text{ A.}$$

The principal contributions to the probable error is due to the uncertainty in the atomic weights of the two sulphur isotopes.

Townes, Holden and Merritt (1947a) find that the OCS line lies at 24325 MHertz, in good agreement with Dakin, Coles and Good, and have also observed the second and third rotational lines respectively of ClCN and BrCN. The frequencies of these lines, and the moments of inertia calculated from them are given in the following table 3:

Table 3

Molecule	Frequency (Hertz)	Moment of inertia (gm/cm^2)
$Br^{79}CN$	24713	$203{\cdot}64 \cdot 10^{-40}$
$Br^{81}CN$	24570	$204{\cdot}83 \cdot 10^{-40}$
$Cl^{35}CN$	23885	$140{\cdot}47 \cdot 10^{-40}$
$Cl^{37}CN$	23389	$143{\cdot}45 \cdot 10^{-40}$

The internuclear distances can be calculated from the isotopic shifts. The values found are:

Bond	Distance between atoms	
	Spectroscopic	Electron diffraction
C–N	1·15 A.	$1{\cdot}13 \pm 0{\cdot}04$ A.
Br–C	1·79	$1{\cdot}79 \pm 0{\cdot}02$
Cl–C	1·64	$1{\cdot}67 \pm 0{\cdot}02$

A fine structure is observed on these lines, which is attributed to nuclear quadrupole moments (see Townes, Holden, Bardeen and Merritt 1947). The agreement with theory is exact if the nuclear spins are assumed to be 3/2 for bromine and 3/2 for chorine; the former agrees with the previous determination by Tolansky (1936), but the latter conflicts with the value of 5/2 for chlorine found by Elliot (1930) and Shrader (1943).

Of non-linear molecules, CH_3I has been examined and the $J=1$ to $J=2$ rotational transition observed near 30000 MHertz by Gordy, Smith and Simmons (1947a). A considerable fine structure is ascribed partly to the effect of a quadrupole moment in the iodine nucleus, and partly to molecules in an excited vibrational state; the analysis is, however, not completed.

Even apart from such fine structures, the analysis of the microwave spectra of the less simple molecules seems likely to offer considerable difficulty, since the spectral lines may be due to transitions between rather high rotational levels which happen to lie close together. The water vapour line at 0·742 cm^{-1} arises in this way, and the analysis of King, Hainer and Cross (1947) shows that even relatively simple molecules such as D_2O, HDO, H_2S and H_2Se should have extensive microwave spectra. Of other molecules, methyl alcohol has been investigated by Dailey (1947) and by Hershberger and Turkevitch (1947) and found to possess a number of absorption lines between 20000 and 27000 MHertz, the interpretation of which is not yet clear. Ten lines in the spectrum of methylamine have been observed in the same region by the latter workers, and there is evidence that this spectrum, and also that of acetone, extends down to 10000 MHertz (Bleaney, Loubser and Penrose, unpublished).

[*Note added in proof*. Recent work on various molecules includes:

CH_3I, Gordy, Smith and Simmons (1947b); CH_3Cl, CH_3Br, Gordy, Simmons and Smith (1947); BrCN, ICN, Gordy, Smith, Smith and Ring

(1947); $O^{16}C^{12}S^{33}$, $O^{16}C^{13}S^{32}$, $Cl^{35}C^{13}N^{14}$, Townes, Holden and Merritt (1947 b); SO_2, Dailey, Golden and Wilson (1947); N_2O, Coles, Elyash and Gorman (1947); H_2O, Golden, Wentink, Hillger and Strandberg (1948); HDO, Strandberg, Wentink, Hillger, Wannier and Deutsch (1948).]

The application of centimetre-wave spectroscopy to atomic spectra has been reported briefly by Roberts, Beers and Hill (1946). By means of the anomalous dispersion near resonance they have detected the weak field Zeeman effect of the hyperfine structure of the ground state of the 133 isotope of caesium, which has a transition frequency of 9192·6 MHertz. The frequency of an oscillator is controlled by means of a resonant cavity (Pound 1946) into which caesium vapour at a pressure of 4.10^{-2} mm. Hg is introduced. On modulating the external magnetic field at 30 c/s. so that a component line of the Zeeman pattern is caused to cross the resonant frequency of the cavity, the anomalous dispersion produces a frequency modulation of the oscillator. The fourteen lines of the Zeeman pattern have been observed by this method.

§6 EXPERIMENTAL WORK ON PRESSURE BROADENING

Apart from the measurements of line breadth already discussed in this report in the section on spectra, an extensive study of pressure broadening of the ammonia spectrum has been made by Bleaney and Penrose (1947 b, 1948 a, b). As mentioned in §5.3, the line breadth constants at a pressure of 0·5 mm. Hg were determined for seventeen lines. These are found to vary in a regular manner, which can be expressed, within the experimental error of 5%, by the relation

$$\Delta\bar{\nu}_{(JK)} = 0{\cdot}50\,.\,10^{-3}(K^2/(J^2+J))^{\frac{1}{2}} \quad (\text{cm}^{-1}). \qquad \ldots\ldots(12)$$

The collision diameters calculated from these measurements lie between 9 A. and 14 A. while the kinetic theory diameter is only 3 A. These large "optical" cross-sections are attributed to dipole–dipole interaction, and the variation of the line breadth constant for the different lines can be explained if it is assumed that a "collision" takes place when the energy of interaction of the dipoles reaches a certain value. Owing to the rapid rotation of the molecule, only that component of the dipole moment which is stationary in space is effective. This component is $\mu K/\sqrt{(J^2+J)}$, and since the interaction energy varies inversely as the cube of the distance between the molecules, the cross-sectional area for a collision varies as $\{K/\sqrt{(J^2+J)}\}^{\frac{2}{3}}$, in agreement with the observed variation of $\Delta\bar{\nu}$. By averaging over all types of collision, the mean energy of interaction required for a collision is found to be (in wave-number units) 1·5 cm^{-1}, which is twice as great as the separation between the two energy levels of the transition. It is thus not surprising that the absorption of radiation can be interrupted by the approach of another ammonia molecule to such a comparatively large distance.

From the measured line breadths at 0·5 mm. Hg pressure (together with values for the remaining lines interpolated by eq. (12)), and the location of the lines in the spectrum, the shape of the absorption curve can be calculated at high pressures. The assumptions which must be made are that the line breadths vary accurately as the pressure, and that no shift in the resonant frequencies occurs. These calculations have been carried out by Bleaney and

Penrose (1947 b) for comparison with their measured values at pressures up to 60 cm. Hg. The calculated and measured absorption curves agree closely at a pressure of 10 cm. Hg, but at 60 cm. Hg the experimental curve shows a definite shift of the absorption towards lower frequencies, though the intensity is of the same order as that calculated. This indicates that the linear variation of line breadth constant with pressure holds accurately from 0·5 mm. Hg to 10 cm. Hg, and probably to 60 cm. Hg. The shift in the absorption curve at 60 cm. Hg is in contrast with the measurements of water vapour, where the centre of the line measured at a pressure of 0·1 mm. Hg of pure water vapour by Townes and Merritt (0·7418 cm^{-1}) is in close agreement with the value of 0·744 cm^{-1} found for water vapour in the atmosphere by Becker and Autler. The width of the line in the latter case is, however, only 0·087 cm^{-1}, whereas at 60 cm. Hg pressure the extrapolated values for the ammonia lines lie between 0·2 and 0·55 cm^{-1}. At 10 cm. Hg pressure the broadest (and most intense) of the ammonia lines have widths of nearly 0·1 cm^{-1}, and there is no observable shift of the absorption. It appears, therefore, that the shift only becomes appreciable when the line breadth constant is nearly as great as the resonant frequency of the lines.

The shape of the absorption curve in ammonia has been studied at pressures up to six atmospheres by Bleaney and Loubser (1948), using wavelengths from 0·8 cm. up to 10 cm. It is found that at pressures above 10 cm. Hg the shape of the absorption curve can be adequately represented by the formulae (5 *b*), (6 *b*) [see § 4] of Van Vleck and Weisskopf, if appropriate values of $\bar{\nu}_{ij}$ and $\Delta\nu$ are used. The inversion spectrum is thus treated as if it were a single line, an approximation which is justified at pressures where the widths of the individual lines are large compared with their separation. The variation of $\bar{\nu}_{ij}$ and $\Delta\bar{\nu}$ with pressure is very interesting. The value of $\bar{\nu}_{ij}$ shifts steadily downwards from 0·78 cm^{-1} (zero pressure) as the pressure rises, becoming substantially zero at pressures above two atmospheres. The value of $\Delta\bar{\nu}$ does not increase linearly with the pressure but can be represented by the formula

$$\Delta\bar{\nu}/p = 0{\cdot}62(1 - 0{\cdot}25p)\ cm^{-1}\ (p \text{ in atm.})$$

up to about two atmospheres. The decrease in $\Delta\bar{\nu}/p$ with increasing p corresponds to the effects of double collisions, since a molecule spends a substantial fraction of its lifetime in "collision", and the impact of a third molecule during this time cannot cause any further interruption of the absorption of radiation. Above two atmospheres $\Delta\bar{\nu}/p$ tends again to a constant value, corresponding to a collision diameter of 7·6 A., which is considerably smaller than that found at a pressure of 0·5 mm. Hg for the individual inversion lines (9–14 A.). Since at these pressures $\bar{\nu}_{ij} = 0$, the high-frequency inversion of the ammonia molecule must be destroyed by the powerful electric fields of neighbouring molecules, and the absorption is due to a different mechanism such as the low-frequency inversion observed in heavy molecules (§ 5.4).

At low pressures where the width of spectral lines is narrow compared with their separation, the intensity at the centre of a line should be independent of the pressure so long as broadening due to causes other than collisions is negligible. Bleaney and Penrose (1948 a) find, however, that in a resonant cavity the

intensities at the centres of lines in the ammonia spectrum begin to fall at pressures of about 10^{-1} mm. Hg, whereas the only other important form of broadening, the Doppler effect, should not become appreciable until pressures of about 10^{-3} mm. Hg are reached. The effect is found to be due to the disturbance of thermal equilibrium caused by the absorption of energy from the radiation. A theory of the effect is developed in which the fall in the absorption coefficient is related to the energy density of the radiation, the transition probability, the line width and the thermal relaxation time. Each of these quantities, except the last, is known from theory or measurement, and thus a comparison with the observed fall in the absorption at low pressures yields a value of the thermal relaxation time. The results of these experiments are shown in figure 4, where the variation in the absorption coefficients at the centre of the line $(J, K) = (3, 3)$ is shown as a function of the pressure for two different levels of energy density in the resonator. The experimental points agree closely with the continuous lines representing the calculated absorption for a thermal relaxation time equal to $1{\cdot}7_5$ times the mean time between collisions which are effective in interrupting the absorption of radiation. This shows that a closer approach between the ammonia molecules is required for a collision which is effective in re-establishing thermal equilibrium than for a collision which only disturbs the absorption, though thereby contributing to the line breadth.

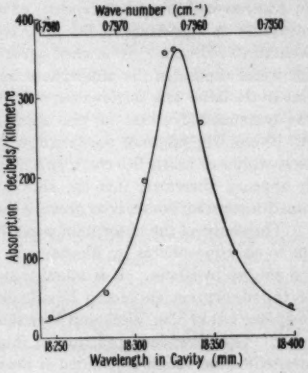


Figure 4.

This phenomenon of the disturbance of thermal equilibrium may also be observed in the waveguide transmission method, though the energy density is smaller for a given power than in the resonant cavity. Bleaney and Penrose (1946 b) have suggested that the low intensities observed by Good for the ammonia lines may be due to this effect; independently, Townes (1946) has shown that the observed diminution in the absorption coefficient is consistent with the disturbance of thermal equilibrium, if the thermal relaxation time is approximately equal to the mean time between collisions which interrupt the absorption. Other observations of the saturation effect have been reported by Smith and Carter (1947) and Pond and Cannon (1947).

Measurements of the collision cross-section for mixtures of ammonia with various non-polar gases have been made by Bleaney and Penrose (1948 b). By observing the ratio of the intensity at the centre of the line (3, 3) for pure ammonia (at pressures up to 1 mm. Hg) to that for a known mixture, the relative cross-section areas for ammonia–ammonia and ammonia–foreign gas collisions can be found, allowance being made for the different molecular velocities.

The observed values are compared with values calculated from an extension of the simple theory of the width due to self-broadening (table 4).

Table 4. Ratio of cross-sectional area (σ_{12}) for broadening by a non-polar gas, to that for self-broadening (σ_{11})

Gas	σ_2/σ_{11} (calc.)	σ_{12}/σ_{11} (observed)	Collision diameter observed (in A.)
Helium	0·067	0·030	2·4
Hydrogen	0·106	0·064	3·5
Oxygen	0·129	0·123	4·9
Nitrogen	0·135	0·21	6·4
Argon	0·135	0·11	4·7
Carbon disulphide	0·245	0·29	7·6
(Ammonia)	(1·00)	(1·00)	(14)

The observed values are in rough agreement with those calculated; they are generally rather smaller for the more rapidly moving molecules. It will be seen that the cross-sections for collisions with non-polar gases are much smaller than for collisions with other ammonia molecules. The same phenomenon has been found for water vapour; there is thus no doubt that the large cross-sections observed for self-broadening are due to dipole–dipole interaction.

§7. CONCLUSION

The use of radio frequencies for the investigation of the fine structure of energy levels, hardly begun ten years ago, has developed to an enormous extent. The present article has been limited to molecular spectra, observed at wavelengths of about one centimetre. It is inevitable, however, that in a field of research advancing so rapidly, no survey can be up to date by the time it appears in print. The author is particularly conscious that insufficient space has been devoted to effects connected with the electric quadrupole of the nucleus, which will probably form one of the most important features of this work. As yet, however, no full paper has appeared on this subject, and there are minor discrepancies in nomenclature between different workers (see Feld 1947). It seems justifiable, therefore, to leave a full survey until more theoretical work has been published.

REFERENCES

Adel, 1947. *Phys. Rev.*, **71**, 806.
Becker and Autler, 1946. *Phys. Rev.*, **70**, 300.
Beringer, 1946. *Phys. Rev.*, **70**, 53.
Bleaney and Loubser, 1948. *Nature, Lond.* (in press).
Bleaney, Loubser and Penrose, 1947. *Proc. Phys. Soc.*, **59**, 185.
Bleaney and Penrose, 1946 a. *Nature, Lond.*, **157**, 339.
Bleaney and Penrose, 1946 b. *Phys. Rev.*, **70**, 775.
Bleaney and Penrose, 1947 a. *Proc. Roy. Soc.*, A, **189**, 358.
Bleaney and Penrose, 1947 b. *Proc. Phys. Soc.*, **59**, 418.
Bleaney and Penrose, 1948 a, b. *Proc. Phys. Soc.*, **60**, 83.
Cleeton and Williams, 1934. *Phys. Rev.*, **45**, 234.
Coles, Elyash and Gorman, 1947. *Phys. Rev.*, **72**, 973.
Coles and Good, 1946. *Phys. Rev.*, **70**, 979.
Dailey, 1947. *Phys. Rev.*, **72**, 84.
Dailey, Golden and Wilson, 1947. *Phys. Rev.*, **72**, 871.

DAILEY, KYHL, STRANDBERG, VAN VLECK and WILSON, 1946. *Phys. Rev.*, **70**, 984.
DAKIN, COLES and GOOD, 1946. *Phys. Rev.*, **70**, 560.
DAKIN, GOOD and COLES, 1947. *Phys. Rev.*, **71**, 640.
DEBYE, 1929. *Polar Molecules* (New York : Chemical Catalog Co.).
DENNISON, 1931. *Rev. Mod. Phys.*, **3**, 280.
DENNISON, 1940. *Rev. Mod. Phys.*, **12**, 189.
DICKE, 1946. *Rev. Sci. Instrum.*, **17**, 268.
DICKE, BERINGER, KYHL and VANE, 1946. *Phys. Rev.*, **70**, 340.
DIEKE and BABCOCK, 19327. *Proc. Nat. Acad. Sci.*, **13**, 670.
ELLIOT, 1930. *Proc. Roy. Soc.*, A, **127**, 638.
FELD, 1947. *Phys. Rev.*, **72**, 1116.
FOLEY, 1947. *Phys. Rev.*, **72**, 504.
FRÖHLICH, 1946. *Nature, Lond.*, **157**, 478.
GOLDEN, WENTINK, HILLGER and STRANDBERG, 1947. *Phys. Rev.* (in press).
GOOD, 1946. *Phys. Rev.*, **70**, 213.
GOOD and COLES, 1947. *Phys. Rev.*, **71**, 383.
GORDY and KESSLER, 1947 a, b. *Phys. Rev.*, **71**, 640 ; **72**, 644.
GORDY, SIMMONS and SMITH, 1947. *Phys. Rev.*, **72**, 344.
GORDY, SMITH and SIMMONS, 1947 a, b. *Phys. Rev.*, **71**, 917 ; **72**, 249.
GORDY, SMITH, SMITH and RING, 1947. *Phys. Rev.*, **72**, 259.
HADLEY and DENNISON, 1946. *Phys. Rev.*, **70**, 780.
HAINER, KING and CROSS, 1946. *Phys. Rev.*, **70**, 108.
HERSHBERGER, 1946. *J. Appl. Phys.*, **17**, 485.
HERSHBERGER and TURKEVITCH, 1947. *Phys. Rev.*, **71**, 554.
HUGHES and WILSON, 1947. *Phys. Rev.*, **71**, 562.
JAUCH, 1947. *Phys. Rev.*, **72**, 715.
JEN, 1947. *Phys. Rev.*, **72**, 986.
KING, HAINER and CROSS, 1947. *Phys. Rev.*, **71**, 433.
LAMB, 1946. *Phys. Rev.*, **70**, 308.
LORENTZ, 1916. *The Theory of Electrons* (New York : Stechert & Co.).
NIELSEN and DENNISON, 1947. *Phys. Rev.*, **72**, 1101.
POLDER, 1942. *Physica*, **9**, 908.
POND and CANNON, 1947. *Phys. Rev.*, **72**, 1121.
POUND, 1946. *Rev. Sci. Instrum.*, **17**, 490.
RANDALL, DENNISON, GINSBURG and WEBER, 1937. *Phys. Rev.*, **52**, 160.
ROBERTS, BEERS and HILL, 1946. *Phys. Rev.*, **70**, 112.
SAXTON, 1947. *Meteorological Factors in Radio-Wave Propagation* (Physical Society), p. 215.
SCHLAPP, 1937. *Phys. Rev.*, **51**, 343.
SHENG, BARKER and DENNISON, 1941. *Phys. Rev.*, **60**, 786.
SHRADER, 1943. *Phys. Rev.*, **64**, 57.
SMITH and CARTER, 1947. *Phys. Rev.*, **72**, 638.
STRANDBERG, KYHL, WENTINK and HILLGER, 1947 a, b. *Phys. Rev.*, **71**, 326 ; **71**, 639.
STRANDBERG, WENTINK, HILLGER, WANNIER and DEUTSCH, 1948. *Phys. Rev.* (in press).
SUTHERLAND, LEE and WU, 1939. *Trans. Faraday Soc.*, **35**, 1373.
TOLANSKY, 1936. *Proc. Roy. Soc.*, A, **136**, 585.
TOWNES, 1946. *Phys. Rev.*, **70**, 665.
TOWNES, 1947. *Phys. Rev.*, **71**, 909.
TOWNES, HOLDEN and MERRITT, 1947 a, b. *Phys. Rev.*, **71**, 64 ; **72**, 513.
VAN VLECK, 1947 a, b. *Phys. Rev.*, **71**, 413 ; **71**, 425.
VAN VLECK and WEISSKOPF, 1945. *Rev. Mod. Phys.*, **17**, 227.
WALTER and HERSHBERGER, 1946. *J. Appl. Phys.*, **17**, 814.
WATTS and WILLIAMS, 1947. *Phys, Rev.*, **71**, 639.
WATTS and WILLIAMS, 1947 a, b, c. *Phys. Rev.*, **72**, 263 ; **72**, 980 ; **72**, 1122.
WILLIAMS, 1947. *Phys. Rev.*, **72**, 974.

ULTRASONICS RESEARCH AND THE PROPERTIES OF MATTER

By CHARLES KITTEL,*
Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts
Now at Bell Telephone Laboratories, Murray Hill, New Jersey.

§ 1. INTRODUCTION

THE principal aim of this Report is to consider the application of ultrasonic methods to the study of the gaseous, liquid, and solid states of matter. It is proposed also to discuss the major contribution to ultrasonic instrumentation made by the electronic techniques associated with the development of radar.

The propagation of sound waves in gases, liquids and solids is considered in §§ 2, 3, and 4, respectively. Experimental methods are discussed in § 5, and the applications to other fields in § 6. The unusual phenomena associated with sound propagation in liquid helium are discussed in § 7. The derivations of the basic equations describing the physical processes involved in sound propagation are widely scattered in the literature; elementary proofs of the major relations are collected together in this Report, particularly in §§ 2 and 3.

The bibliography at the end is intended to be representative of the papers published in ultrasonics since the beginning of 1942. An effort has been made to include, in addition, references to some of the classical papers in the field and also to the more important papers published in enemy and enemy-occupied countries since the beginning of 1940. Owing to wartime conditions, many of the papers cited have not been available to the author. A number of wartime developments of lasting interest are at present described only in laboratory reports which have been given little public circulation. Questions of prior discovery in such cases are often hopelessly tangled, and the author apologizes here for all errors and omissions on this score.

A word on nomenclature is relevant. Present usage tends to adopt the adjective *ultrasonic* in preference to *supersonic* which is now widely used with another meaning in aeronautics, as in "supersonic flow". Workers in India particularly have used *hypersonic* to denote ultrasonic phenomena (of thermal origin) at frequencies of the order of 1000 Mc/sec. or higher. This is a convenient demarcation.

The number of papers published in the general field of ultrasonics during the past twenty years is of the order of several thousand. Hiedemann (1939) refers to over 1300 papers in his text-book. It is probable that the fruits of much of the fine experimental work in the field have appeared largely in the practical applications to metallurgy, chemistry and biology; the scale of research activity in ultrasonics is probably somewhat out of proportion to the results which have

* John Simon Guggenheim Memorial Foundation Fellow

been thus far achieved in terms of the increase of knowledge of the fundamental physical properties of matter.

A few numbers may be useful for a general orientation in the subject. The primary quantities in which one is interested are the velocity and the absorption coefficient of sound. Sound velocities are commonly of the order of 250 m/sec. in gases, 1250 m/sec. in liquids, and 5000 m/sec. in solids. The dependence of the velocity on frequency is usually small. The absorption coefficient is very frequency-dependent; values of the absorption are such that propagation over a distance of the order of 10 cm. may be difficult in many cases when the frequency exceeds 1 Mc/sec. in gases and 50 Mc/sec. in liquids and solids. At these frequencies the wavelengths are of the order of 0·03 cm. in gases, 0·003 cm. in liquids, and 0·010 cm. in solids. Individual substances may show large deviations from the representative values given here, particularly with respect to absorption. The sensitivity of the absorption to small structural changes makes it on the whole a more interesting quantity to study than the velocity.

It is appropriate to consider briefly the interpretation of the physical quantities determined by ultrasonic methods in relation to the basic limitations imposed by the nature of the measurements and the available experimental technique. Detailed discussions of the various topics are given in the separate sections of this Report.

(*a*) *Gases.* The velocity of sound in a *monatomic* gas is determined completely by the equation of state and by the knowledge that the adiabatic condition is approximately satisfied. Sound-velocity measurements in monatomic gases give no information which is not also accessible by means of static p, V, T measurements. Values of the attenuation of sound in monatomic gases can in principle be correlated directly with static determinations of viscosity and thermal conductivity. The sound velocity is independent of frequency; a small dispersion is expected theoretically at high frequencies, but only when the attenuation is so great as to render the effect practically unobservable.

In *polyatomic* gases the adiabatic condition may be modified through the effect of certain time constants which characterize the rate of exchange of thermal energy between the translational, rotational and vibrational states of the molecules. The sound velocity and more particularly the attenuation are functions of the facility of energy exchange at the frequency of the sound wave. The phenomenon is known as thermal relaxation.

The thermal relaxation process is probably the most important physical discovery that has come out of ultrasonics research. However, the present understanding of the subject is largely phenomenological. The relevant relaxation frequencies are difficult to calculate theoretically and to measure experimentally. The experimental difficulty is caused largely by the sensitivity of the observed relaxation frequencies to the presence of minute traces of impurity molecules.

(*b*) *Liquids.* In a monatomic liquid the velocity of sound is determined through the compressibility by the equation of state and the adiabatic condition. The attenuation is directly related to the static viscosity and thermal conductivity.

In polyatomic liquids relaxation processes are important, just as in polyatomic gases. Impurity effects, so important in gases, play a minor rôle here.

However, the experimental knowledge of the relaxation process in liquids is very limited, since for most liquids the relaxation frequencies are in the region of 10^3 to 10^5 Mc/sec., which is above the upper limit (near 10^3 Mc/sec.) of existing absorption measurements. On the theoretical side, practically no progress has been made with the interpretation of relaxation phenomena in liquids because of the rudimentary condition of the theory of the liquid state.

It should be emphasized that in most liquids, as in gases, the order of magnitude of the absorption is determined almost entirely by the relative proximity of a relaxation frequency; that is, a high absorption indicates that a relaxation frequency is nearby, while a low absorption indicates the remoteness of the relaxation frequencies. *The relaxation frequencies are the most significant data to come out of ultrasonic measurements.*

(*c*) *Solids.* In general the velocity of sound in crystalline solids is determined by the static values of the elastic constants and by the adiabatic condition. The velocity is expected to remain constant until the wavelength becomes comparable with the size of the unit cell of crystals, which occurs at frequencies of the order of 10^2 Mc/sec. In plastic materials, however, rela ation dispe sion has been observed in the neighbourhood of 0·2 Mc/sec.

The absorption of sound in solids has been little studied because of the difficulty and the inadequacy of the experimental methods. Now, however, the pulse technique is available, and this field may be expected to develop rapidly. Using the older methods in the range 0·1 to 100 l c./sec. useful work has been done on the study of internal friction in solids. Acoustic wavelengths equal to common grain dimensions in polycrystalline metal specimens correspond to frequencies in the range 1 to 1000 Mc/sec. It appears likely that absorption measurements will be of practical importance as a convenient method for determining grain sizes.

§ 2. PROPAGATION IN GASES*

(i) *Introduction*

The classical equations for sound absorption in gases arising from shear viscosity (Stokes, 1841) and heat condition (Kirchhoff, 1868) are well known. It is proposed here to develop the classical equations in a manner which is easily extended to include the phenomenological theory of thermal relaxation (Herzfeld and Rice, 1928) and other effects of interest.

The rather trivial example of the first-order sound propagation equation will be considered first in order to illustrate a technique of analysis that will prove particularly convenient for the subsequent discussion of the absorption effects.

(ii) *Sound velocity in an ideal gas*

The equation of a plane sound wave propagated in the direction of the positive x axis may be written

$$P = P_0 e^{j(\omega t - kx)}; \quad [\omega = 2\pi f; \ k = 2\pi/\lambda], \qquad \ldots\ldots(1)$$

where P is the local excess pressure in the gas, ω is 2π times the frequency f, and k (if real) is 2π divided by the wavelength λ. It is desired to find the velocity of propagation $c = f\lambda = \omega/k$ in terms of the physical properties of the medium.

* A list of symbols used is given in Appendix I to this Report.

At a given frequency the problem reduces to the determination of the "wave vector" k.

The wave motion is determined by the equation of state

$$pV = \mathrm{n}RT, \text{ or } p = (RT/M)\rho, \qquad \ldots\ldots(2)$$

where M is the molecular weight and ρ is the density; the force equation

$$\frac{\partial u}{\partial t} = -\frac{1}{\rho}\frac{\partial p}{\partial x} = -\frac{RT}{M}\left(\frac{\partial \ln T}{\partial x} + \frac{\partial \ln \rho}{\partial x}\right), \qquad \ldots\ldots(3)$$

where u is the velocity of a volume element; the continuity equation

$$(\partial\rho/\partial t) + (\partial \rho u/\partial x) = 0; \qquad \ldots\ldots(4)$$

the first law of thermodynamics $\Delta U + p\Delta V = \Delta Q$, which may be written

$$C_V(\partial T/\partial t) - (pM/\rho^2)(\partial\rho/\partial t) = 0, \qquad \ldots\ldots(5)$$

since $\Delta Q = 0$ for an adiabatic process, and $p\Delta V = -(pM/\rho^2)\Delta\rho$ from equation (2).

For u, s and θ of the form $e^{i(\omega t - kx)}$, the equations reduce to a set of algebraic equations. Here s and θ are defined by the relations $\rho = \rho_0(1+s)$ and $T = T_0(1+\theta)$. One has, neglecting squares and cross-products of s, u and θ,

$$\omega u = (RT_0/M)(k\theta + ks) = 0; \qquad \ldots\ldots(3\,a)$$

$$\omega s - ku = 0; \qquad \ldots\ldots(4\,a)$$

$$\omega C_V T_0\theta - (pM/\rho)\omega s = 0. \qquad \ldots\ldots(5\,a)$$

This set of three homogeneous linear equations in u, s and θ has a non-trivial solution only if the determinant formed by the coefficients vanishes. The requirement that the determinant vanish reduces to the equation

$$k^2 = (M/\gamma RT)\omega^2, \qquad \ldots\ldots(6)$$

where $\gamma = (C_v + R)/C_v$ is the ratio of the specific heats at constant pressure and at constant volume.

The wave vector k is seen to be entirely real in this case, so that there is no attenuation. The velocity is given by

$$c = \omega/k = (\gamma RT/M)^{1/2}. \qquad \ldots\ldots(7)$$

The *sound velocity* in a gas is of the same order of magnitude as the *molecular velocity*. The sound velocity is equal to $(\gamma/3)^{1/2}$ times the root-mean-square molecular velocity at the same temperature (Jeans, 1925). For $\gamma = 1{\cdot}4$ one has $(\gamma/3)^{1/2} = 0{\cdot}68$.

(iii) *Losses from heat conduction*

Equation (5) expressing the adiabatic condition neglects the effect of the thermal conductivity of the gas which gives rise to some transfer of thermal energy within the sound wave between successive warm and cool half cycles. The strict adiabatic condition $\Delta Q = 0$ must be modified to take account of heat flow. The heat conduction equation may be put in the form

$$K(\partial^2 T/\partial x^2) = (\rho/M)(\partial Q/\partial t), \qquad \ldots\ldots(8)$$

where Q refers to the thermal energy per mole and K is the thermal conductivity. The modified equation expressing the first law of thermodynamics is

$$C_v(\partial T/\partial t) - (pM/\rho^2)(\partial\rho/\partial t) = (MK/\rho)(\partial^2 T/\partial x^2) \qquad \ldots\ldots(9)$$

or

$$\omega C_v\theta - \omega Rs - j(MK/\rho)k^2\theta = 0. \qquad \ldots\ldots(9\,a)$$

Using equation (9*a*) in place of equation (5*a*) the determinantal equation in k becomes

$$k^2=\omega^2(M/\boldsymbol{R}T)\left\{\frac{C_v-jWk^2}{C_p-jWk^2}\right\}, \qquad \ldots\ldots(10)$$

where W is written for $MK/\rho\omega$.

In general k will be complex. Writing

$$k=k_1-jk_2, \qquad \ldots\ldots(11)$$

the equation of a plane wave takes the form

$$P=P_0e^{-k_2x}e^{j(\omega t-k_1x)}, \qquad \ldots\ldots(12)$$

so that k_1 determines the wavelength and indirectly the velocity, and k_2 determines the pressure attenuation. The symbol α is often used in place of k_2. Except in unusual cases of extreme damping, k_1 is much greater than k_2. Usually one may write

$$k^2=k_1^2-k_2^2-2jk_1k_2\simeq k_1^2-2jk_1k_2. \qquad \ldots\ldots(13)$$

With the usual approximation $Wk^2\ll C_v$, equation (10) becomes

$$k_1^2-2jk_1k_2\simeq\omega^2(M/\boldsymbol{R}T)\left\{\frac{C_v}{C_p}-j\frac{Wk^2}{C_p}\frac{C_p-C_v}{C_p}\right\}, \qquad \ldots\ldots(10\,a)$$

whence the sound velocity is given by

$$c=\omega/k_1=(\gamma\boldsymbol{R}T/M)^{1/2}, \qquad \ldots\ldots(14)$$

which is the usual adiabatic velocity. The attenuation from heat conduction is given by

$$\alpha_{th}=k_2=\frac{2\pi^2}{c^3}\frac{\gamma-1}{c_p}Kf^2, \qquad \ldots\ldots(15)$$

where f is the frequency. In this last equation the specific heat c_p refers to unit mass of the gas.

From kinetic theory

$$K\approx\rho c'\Lambda c_p, \qquad \ldots\ldots(16)$$

where c' is the molecular velocity and Λ is the mean free path. On substituting equation (16) in equation (15) one finds

$$\alpha_{th}\approx\frac{\Lambda}{\lambda}\frac{1}{\lambda}, \qquad \ldots\ldots(17)$$

which shows that the pressure falls by $1/e$ in a number of wavelengths of the order of the ratio of the wavelength to the mean free path of the molecules.

The condition $Wk^2\ll C_v$, under which equation (15) was derived, reduces on using the kinetic theory value of the thermal conductivity to the inequality $\Lambda\ll\lambda$. It is seen that this requirement is something quite fundamental, for the entire treatment of a gas as a continuous medium rests on the same assumption. The mean free path is effectively a lower limit to the wavelength of sound waves in a gas.

Herzfeld (1935), Primakoff (1942), and Tsien and Schamberg (1946) have considered the kinetic theory corrections to the continuum theory which arise when the wavelength becomes comparable with the mean free path. The mean free path in helium at S.T.P. is $2{\cdot}5\times10^{-5}$ cm., so that the limiting frequency is of the order of 2000 Mc/sec. A small decrease in velocity is expected as the

limiting frequency is approached; this effect is known as translational dispersion. It has not been observed.

As a matter of formal interest the other limiting case of equation (10) for which $Wk^2 \gg C_v$ will be considered. In this limit the sound velocity is

$$c = (\boldsymbol{R}T/M)^{1/2}, \qquad \ldots\ldots(18)$$

which is the isothermal or Newtonian velocity. Formally this value would be expected for high values of the thermal conductivity and for high frequencies, as was pointed out by Condon (1933). The absorption in this limit is given by

$$\alpha_{th} = \frac{\rho c}{2}\frac{c_p - c_v}{K}. \qquad \ldots\ldots(19)$$

The application of equations (18) and (19) to liquid helium II has been discussed by Groenewold (1939); his considerations are probably not valid according to present theories regarding the nature of liquid helium II.

(iv) *Losses from viscosity*

It may appear difficult at first sight to understand the origin of the shearing motion which causes the viscous losses in a plane wave. However, a motion in which one face of a cubic volume element is displaced while the perpendicular faces remain stationary is not a simple compression. Such a motion may be considered as compounded from a compression and a shear.

The result of the calculation which is given in detail in Appendix II to this Report is that the effective viscosity coefficient is given by $\chi + \frac{4}{3}\eta$, where χ represents the "compressional viscosity" and η the ordinary shearing viscosity. The compressional term will be discussed below in connection with thermal relaxation; for the present the so-called Stokes assumption will be made, according to which $\chi = 0$.

In this case the force equation for the elementary cube leads immediately to

$$(\partial u/\partial t) - (4\eta/3\rho)(\partial^2 u/\partial x^2) = -(1/\rho)(\partial p/\partial x), \qquad \ldots\ldots(20)$$

which replaces equation (3). Similarly, equation (3 *a*) is replaced by

$$\omega u - jk^2(4\eta/3\rho)u - (\boldsymbol{R}T_0/M)(k\theta + ks) = 0, \qquad \ldots\ldots(21)$$

which leads to the determinantal equation

$$k^2 = \omega^2(M/\gamma \boldsymbol{R}T)\frac{1}{1 + j(4\omega\eta M/3\rho\gamma \boldsymbol{R}T)}. \qquad \ldots\ldots(22)$$

In the limiting case $(4\omega\eta M/3\rho\gamma \boldsymbol{R}T) \ll 1$, the sound velocity is given in the usual adiabatic form

$$c = (\gamma \boldsymbol{R}T/M)^{1/2}, \qquad \ldots\ldots(23)$$

while the pressure attenuation from viscosity alone is given by

$$\alpha_{vis} = \frac{8\pi^2}{3}\frac{\eta}{\rho c^3} f^2. \qquad \ldots\ldots(24)$$

From kinetic theory

$$\eta \approx \rho c' \Lambda, \qquad \ldots\ldots(25)$$

so that

$$\alpha_{vis} \approx \frac{\Lambda}{\lambda}\frac{1}{\lambda}, \qquad \ldots\ldots(26)$$

just as was found in equation (17) for α_{th}.

From equations (15) and (24) one finds

$$\frac{\alpha_{th}}{\alpha_{vis}} = \frac{3}{4}\frac{\gamma-1}{c_p}\frac{K}{\eta}, \qquad \ldots\ldots(27)$$

while from kinetic theory $K/\eta=(9\gamma-5)c_v/4$, so that for a monatomic gas with $\gamma=5/3$, one has $\alpha_{th}/\alpha_{vis}\simeq 0{\cdot}5$.

For high frequencies such that the wavelength is of the order of the mean free path, the wave vector k has equal real and imaginary components.

(v) *Losses from thermal relaxation*

Thermal relaxation losses arise from the incomplete establishment of thermal equilibrium in a system. Energy dissipation results whenever all parts of a system are not at the same temperature. The dissipation becomes particularly pronounced when the period of the heating and cooling cycle is comparable with the time required for heat exchange between the different parts of the system. In polyatomic gases there are time lags of the order of 10^{-5} sec. at S.T.P. in the transfer of energy between the internal vibrational degrees of freedom of the molecules and the external translational degrees of freedom. The time lags produce velocity dispersion and anomalous absorption at frequencies of the order of 100 kc/sec. The absorption from this cause is often very large and may be hundreds of times greater than that caused by the classical processes of viscosity and heat conduction. The velocity is affected because the ratio of the specific heats is involved in the velocity, and this ratio is a function of the completeness of the excitation of the internal degrees of freedom at the given frequency.

A phenomenological description of the process was given by Herzfeld and Rice (1928), who treat the internal and external degrees of freedom as distinct thermodynamic systems. The heat capacity C_i and the temperature $T=T_0(1+\theta')$ refer to the internal motion, while C_e and $T=T_0(1+\theta)$ refer to the external motion.

On this picture the first law of thermodynamics is written

$$C_i(\partial T'/\partial t)+C_e(\partial T/\partial t)-(pM/\rho^2)(\partial\rho/\partial t)=0 \qquad \ldots\ldots(28)$$

in place of equation (5), or as

$$\omega C_i T_0\theta'+\omega C_e T_0\theta-(pM/\rho)\omega s=0 \qquad \ldots\ldots(28a)$$

in place of equation (5*a*).

The transfer of energy between the internal and external systems will be supposed to be characterized by a single time constant τ, such that

$$(\partial T'/\partial t)=(T-T')/\tau \qquad \ldots\ldots(29)$$

or

$$j\omega\theta'=(\theta-\theta')/\tau. \qquad \ldots\ldots(29a)$$

This supplementary condition must be considered in setting up the determinantal equation for k.

The determinantal equation is

$$k^2=\omega^2(M/\boldsymbol{R}T_0)\frac{C_v^0+j\omega\tau C_v^\infty}{C_p^0+j\omega\tau C_p^\infty}, \qquad \ldots\ldots(30)$$

giving

$$\begin{aligned} c_0&=(\gamma^0\boldsymbol{R}T/M)^{1/2} \quad [\omega\to 0];\\ c_\infty&=(\gamma^\infty\boldsymbol{R}T/M)^{1/2} \quad [\omega\to\infty]. \end{aligned} \qquad \ldots\ldots(31)$$

Here $\gamma^0 = C_p^0/C_v^0$ and $\gamma^\infty = C_p^\infty/C_v^\infty$, where $C_v^0 = C_e + C_l$ is the effective heat capacity in the limit of zero frequency, and $C_v^\infty = C_e$ is the effective heat capacity in the limit of infinite frequency. The velocity decreases with increasing frequency.

The pressure attenuation is given by

$$\alpha_{\text{relax}} = \frac{\pi f^2}{c_0}\frac{c}{c_0}\left\{\frac{C_p^0 - C_v^0}{C_p^\infty}\frac{C_l}{C_v^0}\right\}\left\{\frac{f_0}{f^2+f_0^2}\right\} \qquad \text{.....(32)}$$

where

$$2\pi f_0 = C_p^0/\tau C_p^\infty, \qquad \text{.....(33)}$$

and c is the velocity at the frequency f.

The maximum absorption per wavelength occurs when $f = f_0$ and is given approximately by

$$\alpha_{\text{relax}}^{\max}\lambda \cong \frac{\pi}{2}\frac{C_p^0 - C_v^0}{C_p^\infty}\frac{C_l}{C_v^0}. \qquad \text{.....(34)}$$

At low frequencies such that $f \ll f_0$ we have

$$\alpha_{\text{relax}} = f^2(\pi/f_0 c)\frac{C_p^0 - C_v^0}{C_p^\infty}\frac{C_l}{C_v^0}, \qquad \text{.....(35)}$$

so that at low frequencies α/f^2 is constant for the attenuation caused by relaxation, just as for heat conduction and viscosity.

It is possible to make a simple estimate of the *maximum absorption caused by relaxation of an internal mode*. The average value of the energy in a sound wave is of the order of $\Delta p \Delta V$. Now $\Delta p/p = -\Delta V/V + \Delta T/T$ for an ideal gas, where the ΔV term on the right represents "mechanical" effects and the ΔT term represents "thermal" effects of the pressure. For an adiabatic process the ratio of the ΔT term to the ΔV term is $(\gamma - 1)$, where γ is the ratio of the specific heats. Let C_l be the specific heat associated with the relaxing degree of freedom; then C_l/C_v represents the fraction of thermal energy in the mode concerned. Thus

$$(C_l/C_v)(\gamma - 1)/\gamma$$

is an approximate upper limit to the ratio of the part of ΔP which is 90° out of phase with ΔV to the part which is in phase with ΔV; this ratio is also an approximate upper limit to the fraction of the total energy of the sound wave dissipated in one cycle. One should of course find absorptions of this magnitude only in the neighbourhood of the appropriate relaxation frequency.

It is of interest to compare this estimate of the peak absorption with an experimental value. For CO_2 gas $(\gamma - 1)/\gamma \sim 0{\cdot}23$, and $C_l/C_v \sim 0{\cdot}3$, so that about 0·07 of the energy is dissipated per cycle. This means that the intensity is down $1/e$ in about 14 wavelengths. The values for γ and C_v used in this estimate refer to low frequencies; with data for high frequencies the estimated value is 8 wavelengths. Measurements by Fricke (1940) show that for CO_2 at the relaxation frequency of ~20 kc/sec. the intensity is down $1/e$ in about 4 wavelengths, indicating that the order of magnitude estimate is generally correct. Using the more accurate relation, equation (34), one estimates about 6 wavelengths.

An excellent review of the literature on the theory of the relaxation process has been given by Richards (1939). Sivian (1947) has pointed out that Rayleigh

in 1899 was aware of the possible importance of sound absorption caused by molecular relaxation.

Tisza (1942) has shown how the hydrodynamical equations can be modified to take account of thermal relaxation processes by introducing the relaxation compressional viscosity coefficient χ into equation (21) in the following way (Appendix II):

$$(\partial u/\partial t)-(1/\rho)[(4\eta/3)+\chi](\partial u^2/\partial x^2)=-(1/\rho)(\partial p/\partial x), \qquad \ldots\ldots(36)$$

where χ must be chosen so that the determinantal equation gives equation (30) as before. Tisza shows further that the relaxation viscosity coefficient χ in the limit of low frequencies ($f \ll f_0$) can be written in the same form as the ordinary kinetic theory viscosity coefficient

$$\eta \approx \rho c' \Lambda \qquad \ldots\ldots(25)$$

by interpreting Λ as the mean free path relative to the excitation of the vibration in question.

Kantrowitz (1946) has considered the effect of thermal relaxation on gas flow problems arising in connection with gas turbines and wind tunnels. He also discusses an impact method for measuring relaxation times.

Relaxation losses occur also in connection with the transfer of thermal energy between the medium and foreign bodies. Such losses will be particularly prominent when the period of the sound wave is close to the relaxation time constant characterizing the flow of thermal energy between the medium and the foreign body. The importance of considering the heat conduction equation in discussing sound propagation near boundaries was frequently forgotten by the classical writers on the theory of sound. This led in some cases to major errors.

Important relaxation losses may occur at the face of an ultrasonic transmitter (Alleman, 1941) and at reflecting surfaces (Herzfeld, 1938; Alleman, 1939). The effect of dust and moisture particles in air is well known (Pfriem, 1941); similarly gas bubbles immersed in liquids may lead to high losses (Pfriem, 1940a; Saneyosi, 1941). Losses in the air column in a microphone have also been studied (Pfriem, 1940b). In §4 of this Report the relaxation effects arising from inhomogeneous temperature distribution in polycrystals are considered.

(vi) *Relaxation frequencies*

Theoretical estimates of relaxation times for gases have been made by Zener (1931), Roy and Rose (1935), Landau and Teller (1936), and Rhodes (1946). The calculations bear out the fact that the probability of energy exchange between translational and vibrational energy is very small at ordinary temperatures, while the exchange between translational and rotational energy is in general more favourable.

The frequency of collision of a gas molecule at S.T.P. is of the order of 5×10^9 collisions per second. If every collision were effective in exchanging translational and vibrational energy the relevant relaxation frequency would be of the order of 1000 Mc/sec. (since $f\sim1/2\pi\tau$). Fricke (1940) finds relaxation frequencies for pure gaseous CO_2, N_2O, COS, and CS_2 of 20, 153, 287, and 379 kc/sec. respectively, so that the probability for exchange of translational

and vibrational energy for these gases varies between 2×10^{-5} and 4×10^{-4} per collision. The relaxation frequency should not be confused with the vibrational frequency of the molecule itself.

Rotational relaxation has been demonstrated only in hydrogen, where E. Stewart (1946) reports a relaxation frequency of about 10 Mc/sec. at S.T.P. on the basis of velocity dispersion and absorption measurements. The exchange probability is about 2×10^{-2} per collision. Vibrational relaxation is not possible here because the vibrational specific heat is not actuated at room temperature.

On the whole, the theory of the probability of energy exchange is in a primitive condition. It is possible, however, to make a *rough estimate* of the probability by means of a method that depends on several assumptions not satisfied in the present situation, so that the results are probably reliable to only an order of magnitude.

According to the method of impact parameters (Williams, 1933) the probability Π of exciting a state of energy $\hbar\omega_s$ in a collision is given by

$$\Pi = \left|\int_{-\infty}^{\infty} \mathrm{F}(t)e^{i\omega_s t}\,dt\right|^2 |x|^2/\hbar^2, \qquad \ldots\ldots(37)$$

where x is the matrix element of the coordinate x connecting the initial and final stares, and $\mathrm{F}(t)$ is the force acting between the molecules.

For analytical convenience one may make the assumption that

$$\mathrm{F}(t) = (W/a_{\mathrm{H}})e^{-r^2/a_{\mathrm{H}}^2}, \qquad \ldots\ldots(38)$$

where W has the dimensions of energy, and a_{H} is the Bohr radius. For head-on collisions

$$\mathrm{F}(t) = (W/a_{\mathrm{H}})e^{-v^2t^2/a_{\mathrm{H}}^2}, \qquad \ldots\ldots(39)$$

where v is the relative velocity of the molecules.

The Fourier transform of $\mathrm{F}(t)$ is

$$(W/v)(1/\pi)^{1/2}\exp[-(\omega_s a_{\mathrm{H}}/2v)^2],$$

so that

$$\Pi = (W|x|/\hbar v)^2(1/\pi)e^{-\omega_s^2 a_{\mathrm{H}}^2/2v^2} \qquad \ldots\ldots(40)$$

On averaging over a Maxwellian distribution of velocities for the relative coordinate system, one obtains

$$\Pi_{\mathrm{av}} = (W|x|/\hbar)^2(m/2\pi\boldsymbol{k}T)^{3/2}2\cdot\int_0^{\infty}\exp[-(\omega_s^2 a_{\mathrm{H}}^2/2v^2)-(mv^2/4\boldsymbol{k}T)]\,dv. \qquad \ldots\ldots(41)$$

The result after integrating and setting $|x|\approx a_{\mathrm{H}}$ is

$$\Pi_{\mathrm{av}} = (Wa_{\mathrm{H}}/\hbar)^2(m/2\pi\boldsymbol{k}T)\exp[-\omega_s a_{\mathrm{H}}/(2\boldsymbol{k}T/m)^{1/2}]. \qquad \ldots\ldots(42)$$

The argument of the exponential may be written $-\omega\tau$, where $\tau = a_{\mathrm{H}}/(2kT/m)^{1/2}$ is approximately the time required for a molecule to traverse the distance a_{H}. The entire exponential factor varies from about 10^{-5} for CS_2 to 10^{-7} for CO_2, where ω_s is the fundamental vibrational frequency of the molecule. The factor in front of the exponential is probably of the order of 100 for vibrational levels, as is found if W is set equal to $mv^2/2$ for the velocity for which the argument of the exponential under the integral in equation (41) is a maximum.

This very crude estimate gives exchange probabilities of the order of 10^{-3} for CS_2 and 10^{-5} for CO_2. It is reassuring that the calculation gives a value

near the correct order of magnitude and also the correct sequence for the different gases. The sequence is determined chiefly by the *vibrational frequency of the molecule*. The fundamental frequencies of the deformation or lowest frequency mode in the linear molecules CO_2, N_2O, COS and CS_2 are 668, 589, 527, and 397 wave numbers, respectively; the order in this sequence is, as expected from equation (42), the reverse of that of the relaxation frequencies.

(vii) *Recent work in gases*

Among the recent experimental studies of ultrasonic propagation in gases are the following:

Acetaldehyde	Alexander and Lambert (1942).
Air	Mokhtar and Richardson (1945); Pielemeier (1945); Quigley (1945); Schilling *et al.* (1947).
Ammonia gas	van Itterbeek and Lauwers (1946); Keller (1940).
Argon	Keller (1940).
Carbon dioxide	Overbeek and Kendall (1941); Keller (1940); Pielemeier (1943); Pielemeier and Byers (1943).
Carbon monoxide	van Itterbeek and Vandoninck (1943).
Deuterium	van Itterbeek and Vermaelen (1942); van Itterbeek and Vandoninck (1943).
Hydrogen	van Itterbeek and Vermaelen (1942); Stewart, Stewart and Hubbard (1945); Stewart (1946).
Methane	Quigley (1945).
Nitric oxide	Kneser (1941).
Nitrogen	Keller (1940).
Propylene	Telfair (1942).
Various vapours	Matta and Mokhtar (1944); Jatkar and Lakshminarayanan (1946).
Mixtures of A, He, H_2	van Itterbeek and Vandoninck (1946).

In recent work there has been some indication that different vibrational modes of a molecule may have *different relaxation times*. This is not an unexpected conclusion, but it is difficult to disentangle experimental proof from the data. In acetaldehyde, which is a relatively complex molecule, Alexander and Lambert (1942) found evidence for three time constants; they mention that evidence from chemical kinetics also suggests the existence of different modes of vibrational activation of the molecule with different probability of decomposition. Pielemeier (1943) found evidence for two relaxation frequencies in carbon dioxide: a major absorption peak at 17 kc/sec. which is associated with the deformation vibration, and a minor absorption peak at 37 kc/sec. which is associated with a valence vibration. The frequencies refer to a pressure of one atmosphere.

As mentioned previously, evidence of *rotational relaxation* has been found recently for the first time. By measurements of ultrasonic absorption in hydrogen and deuterium, van Itterbeek and Vermaelen (1942) deduced rotational relaxation frequencies of the order of 10 Mc/sec. Conclusive proof of the reality of the rotational relaxation process was given by Stewart, Stewart and Hubbard (1945),

and Stewart (1946), who detected in hydrogen the velocity dispersion expected to be associated with the absorption peak. Keller (1940) suggested the possibility of rotational relaxation in nitrogen and ammonia gas.

Kneser (1941) has attempted to detect a new type of relaxation, that associated with the *electronic spin* specific heat in nitric oxide (NO). The ground state of nitric oxide is $^2\Pi$, with a separation of $\sim$120 cm $^{-1}$ between the $^2\Pi_{3/2}$ and $^2\Pi_{1/2}$ components. The energy corresponding to this separation is comparable with kT at room temperature, so that the spin doublet makes an appreciable contribution to the specific heat (Van Vleck, 1932). Kneser failed to detect an effect at low frequencies and concluded that the associated relaxation time τ at S.T.P. was less than 10^{-6} sec. Earlier measurements by van Itterbeek and Thys (1938a) showed that the observed absorption at 600 kc/sec. agreed closely with the classical value based on viscosity and heat conduction alone.

Kohler (1941) has investigated the kinetic theory of sound absorption in *mixtures of monatomic gases*. He finds that the absorption coefficient in mixtures is increased by three new terms arising from diffusion processes. These terms are all small in mixtures of isotopes or atoms of nearly equal masses.

Telfair (1942) used ultrasonic measurements to determine the heat capacity of propylene.

(viii) *Present status of measurements in gases*

Measurements of the velocity of sound in gases are generally in excellent agreement with the theoretical values deduced from the complete equation of state. The correctness of this statement is shown by the fact that values of the second virial coefficient calculated from sound velocity data are in excellent accord with values based on static measurements.

The status of *absorption measurements* is not so clear. In *monatomic gases* one would expect the absorption to be described completely by the classical processes of viscosity and heat conduction, so that the absorption coefficient should be $\alpha = \alpha_{th} + \alpha_{vis}$ from equations (15) and (24). Extensive measurements in argon at 4·25 Mc/sec. by Keller (1940) confirm the classical formula over a wide range of pressure up to 12 atmospheres. This is in agreement with earlier measurements by van Itterbeek and Mariëns (1937*b*).

However, measurements in helium by van Itterbeek and Thys (1938b), van Itterbeek and Mariëns (1940c), and Curtis (1934), and in neon by van Itterbeek and Thys (1938b) indicate that the absorption in these two gases *exceeds the theoretical value* by a factor of the order of four. Pumper (1939) has measured the absorption in helium as a function of pressure and finds that the theoretical absorption is verified from the slope of the plot of absorption against (1/pressure), but that the observed value of the absorption at any pressure is greater than the theoretical value. Pumper suggests that the cause of the observed behaviour may be traced to irregular local excitation of the crystal oscillator leading to important diffraction losses.

The experimental situation is especially confusing, since the observed absorption is in agreement with the theoretical value in argon. Further experimental work is needed in this field before one would be willing to admit the inadequacy of the present theoretical description.

Absorption processes in *polyatomic gases* are understood in general terms on the basis of the relaxation theory. The theory appears to give an adequate account of the observed absorption peaks and the associated velocity dispersion. Little progress has been made, however, in the quantitative correlation of observed and theoretical relaxation times, or in understanding the energy transfer processes within individual molecules. It seems likely that further theoretical work would be profitable here.

The pronounced sensitivity of the relaxation frequencies to small amounts of *impurity gases* is poorly understood at present. For example, the presence of 0·2% of water vapour raises the relaxation frequency in CO_2 gas at S.T.P. from 20 kc/sec. to 400 kc/sec. (Knudsen and Fricke, 1940). The presence of H_2S impurity on the other hand has relatively little effect. It appears then that $H_2O - CO_2$ collisions are greatly more effective in internal energy transfer than $CO_2 - CO_2$ or $H_2S - CO_2$ collisions. Detailed measurements have been made by van Itterbeek and Mariëns (1940a, b). Pertinent theoretical work has been done by Zener (1931) and Kallmann and London (1929).

No theoretical explanation has been advanced to account for the strong dependence of absorption in oxygen on *magnetic field strength*, as reported by van Itterbeek and Thys (1938a).

Mason (1946) has made a direct study of the *shear viscosity of gases*, using a torsional crystal method. His preliminary results indicated that the shear viscosity in polyatomic gases showed relaxation effects. This does not appear to be explicable on the present picture of absorption processes in gases. A shearing motion does not involve alternate heating and cooling of the gas, so that thermal relaxation effects would not be expected to be apparent. It is possible that the experimental situation was confused by the coupling of longitudinal modes of vibration with the torsional motion.

§3. PROPAGATION IN LIQUIDS

(i) *Introduction*

The two traditional concepts of the liquid state regard a liquid as either a very dense gas or else as a solid in which the long-range order between the molecules has broken down. Both of these concepts account qualitatively for the two distinguishing properties of liquids; namely, that a liquid combines the small shear modulus characteristic of gases with the small compressibility characteristic of solids. To explain in a convenient and natural manner the origin of other physical properties of liquids, such as fusion, viscosity, dielectric constant, density and solution effects, sometimes the gas and sometimes the solid point of view is used.

In understanding the propagation of sound in liquids the gas model is particularly suitable. It turns out that sound velocities can be predicted qualitatively from this model; and, further, the absorption of sound in liquids can be related to the absorption processes in gases, which are fairly well understood.

(ii) *Sound velocity*

The velocity of sound in a liquid is commonly of the order of five to ten times the velocity in the vapour at the same temperature. For example, the velocity

of sound in liquid oxygen (Liepmann, 1936) at 90° K. is 910 m/sec., while the velocity in oxygen vapour (van Itterbeek and Mariëns, 1937 a) at the same temperature is 180 m/sec.

The relationship between the velocity of sound in the liquid and gaseous states has been made plausible by the work of Eyring and his collaborators (Eyring and Hirschfelder, 1937; Kincaid and Eyring, 1937, 1938; Hirschfelder, Stevenson and Eyring, 1937), who have given a simple and useful explanation of the relation in terms of the gas model. The relation is derived by Eyring in a qualitative way and is not quite correct; later on, a rigorous equation of state for this model will be considered from which the sound velocity can be calculated unambiguously.

Eyring pictures the sound wa e as travel ing with infinite velocity within a molecule and with gas kinetic velocities through the space between the molecules of the liquid (figure 1). This space is termed the "available volume".

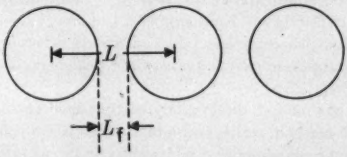


Figure 1. Intermolecular separation L and free-length L_f on Eyring model.
(Reproduced, by kind permission, from the 'Journal of Chemical Physics'.)

The molecules effectively short-circuit a part of the path of the sound wave. If the intermolecular separation is L and the "free length" L_f, the molecules short-circuit all but the fraction L_f/L of the path. The sound velocity in the liquid is then given by

$$c_{\text{liq}} = (L/L_f)c_{\text{gas}}, \qquad \ldots\ldots(43)$$

where kinetic theory gives

$$c_{\text{gas}}^2 = \gamma_{\text{gas}}(\boldsymbol{R}T/M), \qquad \ldots\ldots(44)$$

with $\gamma_{\text{gas}} = C_p/C_v$, the ratio of specific heats in the gas.

The reasoning leading to equation (43) from the specified properties of the model is not rigorous. It is possible to derive the equation of state for the free-volume model, and this has been done by Tonks (1936) in a paper on the statistical mechanics of assemblies of hard elastic spheres. Rice (1942-46) and Frank (1945) have drawn attention to the importance of Tonks' work in relation to theories of the liquid state and have considered some of the thermodynamic consequences of the model. Recently, the application of the model to sound propagation has been studied (Kittel, 1946) in some detail.

(iii) *The Tonks equation of state*

The Tonks equation of state for a classical three-dimensional gas of hard elastic spheres in the limit of close packing is

$$pV(1-\theta^4) = \boldsymbol{N}\boldsymbol{k}T, \qquad \ldots\ldots(45)$$

where θ is the packing fraction with respect to the closest possible packing. One has $\theta = V_0/V = 1-(V_a/V)$, where V_0 is the minimum possible volume and V_a, the available volume, is defined as $V - V_0$. For close packing, as in a liquid,

$$1-\theta^{\frac{1}{3}} \simeq V_a/3V. \qquad \ldots\ldots(46)$$

In this approximation the equation of state becomes

$$pV_a = 3NkT. \qquad \ldots\ldots(47)$$

The form

$$p = (3RT/M)\rho_1, \qquad \ldots\ldots(48)$$

where

$$\frac{1}{\rho_1} = \frac{1}{\rho} - \frac{1}{\rho_0} \qquad \ldots\ldots(49)$$

is also useful. Here ρ_0 is the density corresponding to the closest possible packing.

Before using equation (47), its physical significance as applied to actual liquids will be considered. The pressure p is understood here to be made up of two terms, one representing the effect of the attractive intermolecular forces and the other representing the external applied pressure. Imagine the intermolecular potential energy to be altered by removing the attractive part of the potential while leaving the repulsive part; the "internal pressure" may then be regarded as that pressure which would compress the system to the volume actually occupied. The implied relationship between internal pressure and temperature is plausible on the free-volume model, since the internal pressure does balance the kinetic pressure of the molecules regarded as moving freely within the volume. It is not obvious, however, that the thermal expansion at constant external pressure should be given correctly by the Tonks equation. This amounts to the assumption that the balancing pressure due to potential energy is independent of the volume.

The velocity of sound is given in the usual way by substituting equation (48) in place of the ideal gas equation of state equation (2) and making the corresponding changes in equations (3) to (7), so that

$$c_{\text{liq}} = \left(\frac{V}{V_a}\right)\left(\frac{C_v+3R}{C_v}\right)^{\frac{1}{2}}\left(\frac{3RT}{M}\right)^{\frac{1}{2}} = \left(\frac{V}{V_a}\right)\left(\frac{3\gamma_{\text{liq}}}{\gamma_{\text{gas}}}\right)^{\frac{1}{2}} c_{\text{gas}}. \qquad \ldots\ldots(50)$$

Here γ_{liq} has been set equal to $(C_v+3R)/C_v$, as may be shown to hold as a consequence of equation (47). It may be noted that c_{liq} and c_{gas} are to be taken at the same temperature.

(iv) *Calculation of sound velocity*

It is not practical to give a direct comparison of equation (50) with measurements of sound velocity in liquids, since there is no really accurate independent method for determining the available or the free volume (see, however, Kincaid and Eyring, 1938). One may, however, make an estimate of the available volume by considering the increase in volume between absolute zero and the temperature concerned.

If the free-volume picture is to be taken seriously there should exist some correlation between the molar available volume as calculated, using equation (50)

from sound-velocity measurements at room temperature and the actual increase in molar volume of the material between 0°K. and room temperature. Such a relationship does in fact exist. Let

$$V = V_0 + V_a, \qquad \ldots\ldots(51)$$

where V_0 is the molar volume corresponding to the closest possible packing in a geometrical sense, and V_a is the molar available volume. Let

$$V_a = V_1 + V_2 + V_3, \qquad \ldots\ldots(52)$$

where

V_1 = molar volume change on heating solid from 0°K. to the melting point;
V_2 = molar volume change on fusion; and
V_3 = molar volume change on heating liquid from melting point to the temperature concerned.

A comparison of available volumes as estimated from sound-velocity and thermal expansion data is given in table 1.

Table 1. Comparison of available volume as estimated from thermal expansion and sound-velocity data for 300° K.

	Available-volume increments cm³/mole				Calculated available volume in cm³/mole derived from	
Sub-stance	V_1 0° K. to melting point	V_2 fusion volume change	V_3 melting point to 300° K.	V/V_a from sound velocity	thermal expansion [equation (52)]	sound velocity [equation (50)]
$CHBr_3$	8	10·0	(2)	2·8	20	19
CCl_4	11	4·0	5·0	3·4	20	28
Benzene	10	10·3	2·5	3·5	23	25

From the assumed equation of state one may calculate other quantities of physical interest, including the temperature and pressure coefficients of sound velocity, the thermal expansion coefficient, and the difference of specific heats $(c_p - c_v)$. The results (Kittel, 1946) applied to a number of organic liquids agree in order of magnitude with the experimental results.

It should be noted that the velocity of sound in a liquid depends on the *internal structure* of the constituent molecules only through the effect of the internal degrees of freedom on the specific heat ratio γ. The mechanical internal compressibility of a molecule may be neglected in comparison with the compressibility of the system as a whole. In the case of solutions of high polymers of molecular weight ~100 000, this assumption may, however, require modification.

The internal compressibility of a molecule β_{mole} is of the order of L/q, where L is the length of the molecule and q is the force constant. The force constant is given by $q = (2\pi f)^2 m$, where the frequency f is a fundamental vibration frequency of the molecule and is of the order of 1000 cm^{-1} or 3×10^{13} c.p.s.; m is the reduced mass of the molecule and is of the order of 3×10^{-23} gm. for a light molecule. With these values $q \sim 10^6$ dynes/cm., so that $\beta_{\text{mole}} \sim 10^{-14}$ cm^2/dyne, which

is to be compared with $\beta_{gas} \sim 10^{-10}$ cm^2/dyne under the equivalent pressures obtaining in a liquid.

(v) *Sound absorption*

The three principal mechanisms responsible for the absorption of sound in liquids are the ordinary shearing viscosity, the isothermal compressional viscosity, and the energy loss associated with incomplete excitation of vibrational energy states of the molecules. In many liquids the latter mechanism is the dominant cause of attenuation. The theory developed for gases was first applied to liquids by Mandelstam and Leontovic (1936-37), and later extended by Kneser (1938). Herzfeld (1941) has given an interesting discussion of absorption in liquids, while Tisza (1942) has considered the important implications of the subject with respect to the hydrodynamical equations. Anselm (1945) and Hall (1947) have given a theory of the absorption associated with structural relaxation of the short-range order in liquids.

The absorption α_{th} caused by heat conduction is given by equation (15) which holds for both liquids and gases. The mechanism of heat conduction is essentially the same in non-metallic liquids as in gases (Kincaid and Eyring, 1938). In the relation $K \simeq \rho c' \Lambda c_p$ the mean free path Λ must be replaced by the intermolecular free distance L_f.

Equation (27) for the ratio α_{th}/α_{vis} does not apply to liquids because the actual values of the viscosity of liquids are ten or more times the kinetic theory values. The value of the attenuation ratio for liquids other than liquid metals is quite small, of the order of 1/20 or less, so that the losses caused by heat conduction may usually be neglected. In liquid mercury, however, the ratio $\alpha_{th}/\alpha_{vis} \sim 4$, so that heat conduction is more important than viscosity in this case.

The relaxation absorption α_{relax} given by equations (32) to (35) may be carried over to liquids with appropriate changes in the relaxation frequency. The free volume theory of liquids leads to the same result.

Except in the vicinity of relaxation frequencies the attenuation in liquids is less than that in gases by a factor of the order of 10^{-4}. Using the f^2 law one sees that the classical absorption in gases and liquids is roughly equal when the frequency used is 100 times higher for the liquid than for the gas.

A review of the absorption of sound in various liquids at 15 Mc/sec. shows several regularities with regard to molecular structure.

The three most absorbing liquids of low viscosity are non-polar, namely, carbon disulphide, carbon tetrachloride and benzene, with the following values for the experimental absorption relative to the classical absorption:

Substance	$\alpha_{exper}/\alpha_{classical}$
Carbon disulphide (7Mc/sec.)	~ 600
Carbon tetrachloride	25
Benzene	90

Other non-polar liquids, such as heptane, hexane and toluene, are highly absorbing. This tendency suggests that the coupling of translational and vibrational motion is small in molecules with small electrostatic interactions, giving a long lifetime to the vibrational state and a low relaxation frequency.

Highly associated liquids tend to show low, near-classical absorption at 15 Mc/sec. The alcohols and water are good examples of associated liquids, and they show the following values:

Substance	$\alpha_{exper}/\alpha_{classical}$
Methyl alcohol	3
Ethyl alcohol	1
Propyl alcohol	2
Amyl alcohol	1·5
Butyl alcohol	1
Water	3

This tendency suggests that the formation of molecular groups is conducive to the transfer of vibrational energy and therefore leads to high relaxation frequencies.

(vi) *Viscous waves in liquids*

If a liquid is driven by a crystal oscillator which is cut so that the motion of the crystal is parallel to the face, then a highly-damped viscous motion is set up in the liquid. This motion, which is called a viscous wave, is analogous to the so-called heat waves and eddy current waves, and is characterized by the fact that the wave vector k contains $j^{\frac{1}{2}}$ as a factor.

Let the x direction be parallel to the direction of motion of the crystal and the y direction normal to the crystal face; one has for the equation of motion

$$\rho(\partial u/\partial t) = -\eta(\partial^2 u/\partial y^2) \qquad \ldots\ldots(53)$$

or

$$j\omega\rho = \eta k^2, \qquad \ldots\ldots(54)$$

where η is the viscosity. Now

$$k = (\omega\rho/\eta)^{\frac{1}{2}} j^{\frac{1}{2}} = (\omega\rho/2\eta)^{\frac{1}{2}}(1-j), \qquad \ldots\ldots(55)$$

so that

$$u = e^{(j\omega t - k_0 x)} e^{-k_0 x}, \qquad \ldots\ldots(56)$$

where $k_0 = (\omega\rho/2\eta)^{\frac{1}{2}}$.

The resulting motion is highly damped: the amplitude diminishes by $1/e$ in the distance $x = 1/k_0 = \lambda/2\pi$. A further characteristic of this type of motion is that the phase velocity increases as the square root of the frequency:

$$c_{vis} = \omega/k_0 = (2\eta\omega/\rho)^{\frac{1}{2}}. \qquad \ldots\ldots(57)$$

For $\eta = 10^{-2}$ poises, $\rho = 1$ gm/cm^3, and $\omega = 2\pi f = 2\pi \times 10^5$ rad/sec., one has $c_{vis} = 1{\cdot}1 \times 10^2$ cm/sec. and $\lambda = 1{\cdot}1 \times 10^{-3}$ cm.

This solution is due to Stokes, and early experimental work with viscous waves was done by Reiger (1910). Mason (1946–7) has made accurate measurements of the properties of viscous waves, using an ingenious torsional crystal method. He suggests the possibility of using viscous waves as the basis for a continuous-reading viscosity meter and also for determining the viscosity of small samples of a liquid. Since the penetration of a viscous wave is only a few thousandths of a centimetre for a frequency of 100 kc/sec., a thin film will

produce saturation effects. This development promises to be of particular importance in lubrication technology.

(vii) *Shear waves in liquids*

Poisson (1831) suggested that liquids may possess a shear elasticity for rapid motions as well as a viscosity. In this connection Maxwell (1867) supposed that the motion of a liquid was governed by the equation

$$(\partial u/\partial y)=(S_{xy}/\eta)+(\dot{S}_{xy}/\mu), \qquad \ldots\ldots(58)$$

where S_{xy} is the shear stress and μ is the modulus of shear elasticity. A liquid described by this equation is known as a *Maxwellian* or *relaxing liquid*.

If S_{xy} depends on the time in the form $e^{j\omega t}$, one finds

$$(\partial u/\partial y)=(1+j\omega\eta/\mu)(S_{xy}/\eta), \qquad \ldots\ldots(59)$$

so that the medium is characterized by a mechanical relaxation time defined by

$$\tau=\eta/\mu, \qquad \ldots\ldots(60)$$

with an associated relaxation frequency

$$f_0=\mu/2\pi\eta. \qquad \ldots\ldots(61)$$

The process is termed mechanical relaxation rather than thermal relaxation as discussed for gases because there is no heating or cooling cycle with a shear wave.

At low frequencies ($f\ll f_0$) equation (58) becomes the ordinary viscosity equation; at high frequencies ($f\gg f_0$) the shear elasticity is dominant, and the liquid behaves like a solid. At low frequencies a viscous wave is propagated while at high frequencies one expects to find a shear wave.

For light liquids $\eta\sim10^{-2}$ poises, so that using a shear modulus $\mu\sim10^{10}$ dynes/cm., characteristic of ionic crystals one obtains $f_0\sim10^{11}$ cycles per second, which is far above the experimental range. Shear waves of this frequency may, however, play a rôle in determining the heat capacity of liquids, as has been emphasized by Brillouin (1936) and Lucas (1937).

Mason (1947) pointed out that the shear modulus of rubber-like materials was $\sim10^7$ dynes/cm^2, so that by working with highly viscous polymers one would expect to detect the effects of shear waves at reasonable frequencies. For polymerized castor oil $\eta\sim20$ poises so that $f_0\sim10^7/40\pi\sim10^5$ cycles/sec.; Mason actually found the relaxation frequency near this position. He also reports $\mu\sim10^7$ dynes/cm^2 for polyisobutylene (vistanex) which has a viscosity of about 1100 poises.

Ferry (1942) worked with solutions of high polymers; he found shear moduli of the order of 10^4 to 10^6 dynes/cm^2 for solutions of polystyrene in xylene. The shear motion was detected by stroboscopic observation of the strain double refraction (Ferry, 1941).

The hydrodynamic equations for a Maxwellian liquid with particular reference to wave propagation have been considered by Natanson (1901), Isakovich (1939), and Osida (1941).

(viii) *Hypersonic waves in liquids*

At frequencies such that shear waves can be propagated in liquids the velocity of longitudinal waves will be given by the equation applicable to isotropic solids, namely

$$c=[(\kappa+\tfrac{4}{3}\mu)/\rho]^{\frac{1}{2}} \qquad \ldots\ldots(62)$$

instead of the equation

$$c=[\kappa/\rho]^{\frac{1}{2}} \qquad \ldots\ldots(63)$$

usually applicable to liquids. Here κ is the bulk modulus, and μ is the shear modulus.

Measurements of sound velocity at hypersonic frequencies ($\sim$1000 Mc/sec.) have been made by Raman and Venkateswaran (1939), Venkateswaran (1942), Sunanda Bai (1942) and Rank (1946) using a purely optical method developed by Gross (1930) and Ramm (1936) based on the theory of Brillouin (1922–33).

Brillouin predicted that light scattered by the molecules of a liquid through which a sound wave was passing would be modulated with the frequency of the sound wave. The scattered light will have frequencies $f\pm f_s$, where f_s is the frequency of the sound wave. This may be considered as a special case of the *Raman effect*. The scattered light will be visible at a glancing angle ϕ, such that

$$2\lambda_s \sin\phi=\lambda_l, \qquad \ldots\ldots(64)$$

where λ_s is the acoustic wavelength, and λ_l is the optical wavelength in the medium. This equation is similar to the Bragg condition.

For $\phi=45^\circ$ and $\lambda_l=4\times10^{-5}$ cm., one finds $\lambda_s=2{\cdot}8\times10^{-5}$ cm., which corresponds to a frequency of 5000 Mc/sec. for a velocity of $1{\cdot}4\times10^5$ cm/sec. Such a high frequency has not been obtained by ultrasonic methods. However, the thermal lattice vibrations or *Born-Debye waves* which are always present act as just such a wave. The predicted effect manifests itself as light scattering. For this example the calculated shift in optical wavelength (in *vacuo*) is 0·060 A. which can be measured by optical interferometry. It is seen that the wavelength shift in the light scattered at a given angle ϕ is a function of the sound velocity at the frequency f_s.

Raman and Venkateswaran found $c\cong2500$ m./sec. in glycerine at a frequency of $1{\cdot}53\times10^4$ Mc/sec., as compared with the ultrasonic velocity $c=1910$ m./sec. at 10 Mc/sec. Fitting these values by means of equations (62) and (63), they calculated $\mu=2{\cdot}4\times10^{10}$ dynes/cm^2 for the shear modulus, which may be compared with the bulk modulus $\kappa=4{\cdot}6\times10^{10}$ dynes/cm^2. From equation (60) the mechanical relaxation time constant for shear elasticity is $\tau=3{\cdot}5\times10^{10}$ sec., or $f_0=1/2\pi\tau=500$ Mc/sec., which is far below the frequency at which the measure- was made. It should be emphasized that these vibrations are of thermal origin and were *not* generated by artificial means.

Subsequent measurements suggest that velocity dispersion caused by shear elasticity does not occur below 1000 Mc/sec. in ordinary light liquids. It does, however, occur below this frequency in highly viscous liquids (Mason, 1947; Zachoval, 1939).

(ix) *Recent work in liquids*

Among the recent experimental studies of ultrasonic propagation in liquids are the following:

Series of organic liquids.	Willard (1941, 47 a); Rendall (1943); Pellam and Galt (1946); Rapuano (1947).
Electrolytic solutions.	Rüfer (1942); Corey (1943); Dutta and Ghosh (1944); Rutgers (1946); Fox, Herzfeld and Rock (1946); Beyer and Smith (1946); Cefola, Droz, Frankel, Jones, Maslach and Teeter (1946).
Liquid mixtures.	Prasad (1941–45 a); Giacomini (1942); Derenzi and Giacomini (1942); Willis (1947); Mikhailov and Gourevitch (1946).
Water.	Hsu (1945); Jones and Gale (1946); Teeter (1946); Singh (1945); Fox and Rock (1946); Cefola, Droz, Frankel, Jones, Maslach and Teeter (1946); Labaw and Williams (1947).
Benzene.	Quinn (1946).
Carbon disulfide.	Te-Tchao (1946).
CO_2 near critical point.	Noury (1946).
Dichloroethylene.	Baccaredda and Giacomini (1945).
Viscous liquids.	Hunter (1941); Bazulin (1941).
Conductivity of liquid dielectrics.	Seidl (1940).
Effect of electric and magnetic fields.	Barone and Giacomini (1945).
Separation.	Frei and Schiffer (1947).
Acoustic birefringence.	Zvetkov, Mindlina and Makarov (1946).
Ethylene isomers.	Baccaredda and Giacomini (1946)
Decalin—*cis* and *trans*.	Baccaredda (1946).

The work of Prasad (1942, 45 a, 47) on the *velocity of sound in mixtures* is an interesting attempt to correlate known intermolecular forces in liquids with the change in compressibility on mixing.

Hunter's measurements (1941) on ultrasonic absorption in *highly viscous liquids* at 3·5 Mc/sec. are significant. The absorption is found to be fairly close to the classical value, indicating that the thermal relaxation frequencies are unusually high in viscous liquids by comparison with light non-associated liquids.

Empirical rules relating the velocity of sound in different liquids and in the same liquid at different temperatures have been developed. Lagemann and Corry (1942), Lagemann and Dunbar (1945), and Rama Rao (1946) have had some success in an empirical treatment of the velocity of sound as a "bond" property. There does not appear to be a direct physical basis for such an approach, as the compressibility of a liquid depends much more on the forces between molecules, rather than on the forces between the atoms within an individual molecule. The rule essentially expresses the approximate additivity of the molar volume, and the velocity of sound enters as a small correction term to take account

of thermal expansion in the same way as the surface tension is used in the *parachor*. Lagemann (1944) has discussed sound velocity as an index of molecular association in liquids; further remarks have been made by Weissler (1947).

Suryan (1942) and Arenberg (1945) have pointed out that a rough correlation exists between the *viscosity* and the *sound velocity* in liquids. A general connection between viscosity and sound velocity is plausible on the basis of the "hole" theory of liquids. The work of Batschinski (1913) on vi cosity is relevant here. It may be noted that viscous liquids such as castor oil and glycerine show unusually high sound velocity. Viscous liquids are thought to have few "holes", and the relative absence of holes is known theoretically to favour a high sound velocity.

Sound propagation in *solutions of high polymers* has been studied by Ferry (1942), Beuttenmüller (1942) and Mark (1945). Mark analysed the experiments of Schmid and collaborators who showed that chemical valence bonds in macromolecules such as polystyrene could be broken by means of irradiation at 300 kc/sec. with an intensity of 10 watts/cm^2 for a period of a few hours. The opening of the bond was detected by means of the accompanying irreversible change in the viscosity of the solution. The bond energy is of the order of 2 to 4 ev. per bond. Mark calculated that the frictional force between a macromolecule and the solvent during the passage of the sound wave was of the correct order of magnitude to account for the observed effect. His calculation also indicated that the dissolution should occur only for very large molecules, and this conclusion is in agreement with experiment. In somewhat smaller molecules the weaker van der Waals bonds (0·04 to 0·4 ev.) can be broken by ultrasonic irradiation, leading to reversible viscosity changes. The subject has also been considered by Schmid (1940).

The theory of sound propagation in liquids has been considered by Sakadi (1941) and Anselm (1945). The paper by Anselm is of particular importance as it is the first treatment of the effects on sound propagation of *order* in liquid structure. Gorodetzky (1940) has calculated the scattering of sound near the critical point of mixing of two liquids. Eckhart (1946) has emphasized the underlying unity of the various absorption processes as different aspects of the relaxation process.

(x) *Present status of measurements in liquids*

There is no detailed theory of the compressibility of liquids with which the measurements of sound velocity in liquids may be compared. The free volume theory gives only estimates of the order of magnitude of the velocity. Neither is there a theoretical basis for the calculation of relaxation times in liquids, so that at present one may only compare absorption measurements with theory for the case of monatomic liquids, as here thermal relaxation losses are presumed not to enter.

The only monatomic liquids for which the absorption has been measured are mercury and liquid argon. Bär (1937) finds in mercury $\alpha/f^2 = 6{\cdot}6 \times 10^{-17}$ sec^2/cm.; Rieckmann (1939) finds $6{\cdot}5 \times 10^{-17}$; Ringo *et al.* (1947) find $\sim 6{\cdot}0 \times 10^{-17}$ between 20 and 1000 Mc/sec., while the theoretical classical value of α/f^2 in mercury is $5{\cdot}0 \times 10^{-17}$ sec^2/cm. In liquid argon the theoretical classical value of α/f^2 is 3×10^{-17} sec^2/cm., which is in fair agreement with preliminary unpublished measurements in this laboratory.

It is reassuring that the hydrodynamical equations appear to give a correct description of the absorption in *monatomic* liquids in view of the fact that the absorption in *polyatomic* liquids often greatly exceeds the classical value. The theory has been tested up to 1000 Mc/sec. in the case of mercury.

The famous example of "anomalous" absorption in polyatomic liquids is the case of *carbon disulphide*, for which $\alpha/f^2 = 11\,000 \times 10^{-17}$ sec^2/cm. at about 1 Mc/sec., in comparison with the classical value 5×10^{-17}. The value is spectacularly high (Lucas and Bicquard, 1937) in comparison with the classical value, yet at 1 Mc/sec. the pressure is only down by $1/e$ in 90 wavelengths (experimental); by contrast the relaxation absorption in carbon dioxide gas at 20 kc/sec. produces a pressure attenuation of $1/e$ in 8 wavelengths, in agreement with the theory for this case. The velocity dispersion in carbon disulphide between 0·5 Mc/sec. and 10 Mc/sec. should on the relaxation theory amount to about +1 %. An effect of this order is probably within the limits of error of velocity measurements on this particular liquid. The measurements of Te-Tchao (1946 a) are particularly suspect as he finds a negative dispersion, which is the wrong sign for the relaxation theory. Ginsburg (1942) shows that a small negative dispersion at *hypersonic* frequencies might be explained on the basis of second order correction terms in the hydrodynamic equations. Recent measurements by Rapuano (1947) at 75 and 105 Mc/sec. indicate that $\alpha/f^2 = 3000 \times 10^{-17}$ sec^2/cm. at both frequencies.

It seems then that the behaviour of carbon disulphide is not particularly remarkable. There is no compelling reason at present to believe that the apparent high absorption in the megacycle range is necessarily incompatible with theory.

In order to gain a real understanding of relaxation processes in liquids it is necessary to have absorption measurements on the same liquids covering a very wide frequency range, preferably from 100 kc/sec. to the hypersonic region above 1000 Mc/sec. Such a programme will require an extension to both higher and lower frequencies than are currently used for liquid measurements.

It is clear from the measurements of Fox and Rock (1946) on the temperature dependence of the absorption coefficient for water that the *classical mechanisms and thermal relaxation cannot together account for all of the absorption in water.* The basis for this statement is the fact that the thermal expansion coefficient and hence $(c_p - c_v)$ vanish near 4° C. in water, so that the sound wave proceeds isothermally, without alternate heating and cooling. If the process is isothermal there is no occasion for the population of the internal vibrational states of the molecules to change, so that there can be no relaxation loss caused by a delay in thermodynamic equilibrium.

The measured value of the absorption in water at 4° C. at 17 Mc/sec. is $\alpha/f^2 = 50 \times 10^{-17}$ sec^2/cm., to which the classical terms contribute about 15×10^{-17}; by the above argument the thermal relaxation term must be zero. The unexplained term amounts to about 35×10^{-17}. This term appears to have the form of an isothermal compressional viscosity and may, as suggested by Debye (1939), arise from the mechanical relaxation of the short range order structure of the liquid with a relaxation frequency $\sim 10^5$ Mc/sec.

It is plausible to suppose that the relaxation frequencies for shear and compressional elasticity are comparable, and this assumption leads to an isothermal

compressional viscosity comparable with the shear viscosity. A detailed discussion on the basis of reaction-rate theory has been given by Frenkel (1946) and Hall (1947).

§ 4. PROPAGATION IN SOLIDS

(i) *Introduction*

The differences which exist between sound propagation in liquids and solids arise primarily from the rigidity of solids, and secondarily from their anisotropy. The rigidity makes possible various kinds of wave motion, and the anisotropy leads in particular to two important types of sound absorption.

In an infinite isotropic solid there are two principal types of wave motion, namely *longitudinal* waves and *shear* waves. The velocity of longitudinal waves is given by (appendix II)

$$c_{\text{long}} = [(\kappa + \tfrac{4}{3}\mu)/\rho]^{\frac{1}{2}}, \qquad \ldots\ldots(65)$$

where κ is the bulk modulus, and μ is the shear modulus. The velocity of shear waves is given by

$$c_{\text{shear}} = [\mu/\rho]^{\frac{1}{2}}, \qquad \ldots\ldots(66)$$

which is lower than c_{long} by a factor of the order of one-half. Values of $c_{\text{shear}}/c_{\text{long}}$ for fused quartz, steel, and lead are 0·65, 0·52, and 0·30, respectively. The velocity of longitudinal waves in a wire whose diameter is small in comparison with the wavelength is given by

$$c_{\text{wire}} = [E_S/\rho]^{\frac{1}{2}}, \qquad \ldots\ldots(67)$$

where E_S is the adiabatic Young's modulus. This velocity is about 12% less than c_{long} in the case of steel.

The absorption in single crystals and in amorphous solids is usually less than that in liquids and of course much less than that in gases. In polycrystalline specimens the effect of crystal anisotropy increases the absorption. The elastic anistropy of individual crystallites which are randomly oriented makes the material appear inhomogeneous to applied stress. The inhomogeneities lead to scattering of the Rayleigh λ^{-4} type and also to thermal relaxation losses of the Zener type associated with the inhomogeneous heating and cooling of the material during the passage of a sound wave.

Solids show absorption losses caused by heat conductivity just as in gases and liquids, and also by plastic flow, which can be interpreted as a type of viscosity. There are also several other causes of ultrasonic absorption in solids. Some theoretical or experimental work has been done on the processes listed here.

Mechanism	Occurrence	Approximate Frequency Range	General Reference
Thermoelastic relaxation	Polycrystals	1–100 kc/sec.	Zener (1938)
Scattering	Polycrystals	1 Mc/sec. and up	Mason and McSkimin (1947)
Plastic flow	All solids	All frequencies	Zener (1946)
Thermal conductivity	All solids	All frequencies	—
Structural relaxation	Plastics (?)	50–500 kc/sec.	Mason (1942)
Anharmonic coupling	All solids	Above 1000 Mc/sec.	Landau and Rumer (1937)
Magnetic effects	Ferromagnetics	Up to 1000 Mc/sec.	Becker and Döring (1939)

(ii) *Thermoelastic relaxation*

The nature of the thermoelastic internal friction can be understood from an elementary example discussed by Zener (1937). In a *vibrating reed* the thermoelastic part of the internal friction arises from the flow of heat back and forth across the reed. When the reed is bent the inner side is heated, and the outer side is cooled. With each half cycle of the motion the sign of the temperature differential is reversed.

If the motion takes place very slowly compared with the time required for the heat flow across the reed, the temperature remains constant and the transfer of mechanical energy into heat takes place reversibly. There is therefore no internal friction in the *isothermal* limit. If, on the other hand, the motion takes place so rapidly that no heat flow occurs there is also no internal friction caused by this particular effect. This is the *adiabatic* limit. At intermediate frequencies there is, however, a resultant heat flow which results in loss of energy.

The frequency dependence of the internal friction is typical of relaxation processes. Zener finds the thermoelastic Q of a reed vibrating transversely to be

$$Q^{-1} = \frac{E_S - E_T}{E_S} \cdot \frac{f_0 f}{f_0^2 + f^2}, \qquad \ldots\ldots(68)$$

where E_S is the adiabatic and E_T the isothermal Young's modulus.

In the case of agreed rectangular cross section, the relaxation frequency f_0 is given by

$$f_0 = (\pi/2)(D/d^2), \qquad \ldots\ldots(69)$$

where $D = K/c_p\rho$ is the thermal diffusion constant of the material and d is the thickness of the reed in the plane of vibration. For polycrystalline aluminium at room temperature $K = 0{\cdot}5$ cal/cm² sec. °c., $c_p = 0{\cdot}2$ cal/gm., and $\rho = 2{\cdot}7$ gm/cm³, so that $D \cong 1$ cm²/sec. Taking $d = 10^{-3}$ cm., one finds $f_0 = 1{\cdot}6$ Mc/sec.

Values of Q may be converted to values of the pressure attenuation coefficient α in the following way: Q is defined by

$$Q = 2\pi\left(\frac{\text{energy stored}}{\text{energy dissipated/cycle}}\right). \qquad \ldots\ldots(70)$$

The energy density in a plane sound wave is given by $P^2/2\rho c^2$, and the intensity is $P^2/2\rho c$, where $P = P_0 e^{-\alpha x}$. The rate of energy dissipation in a slab of unit area and thickness Δx in the direction of the sound wave is given by the difference between the intensities into and out of the slab, or by $(\partial I/\partial x)\Delta x$, which is equal to $P^2\alpha\Delta x/\rho c$. The dissipation in the slab per cycle is $P^2\alpha\Delta x/\rho cf$, so that $Q = 2\pi[(P^2\Delta x/2\rho c^2)/(P^2\alpha\Delta x/\rho cf)]$. This gives

$$\alpha = (\pi/\lambda)Q^{-1} = (\pi f/c)Q^{-1}. \qquad \ldots\ldots(71)$$

Calculations of the thermal damping for bodies of various shapes are given by Päsler (1944).

Zener (1938) pointed out that the stress in a *vibrating polycrystalline solid* varies from crystallite to crystallite, because the elastic anisotropy of single crystals combined with random orientation renders it impossible for strains which are macroscopically homogeneous to be also microscopically homogeneous, This stress fluctuation gives rise to heat currents between adjacent crystallites, which in turn gives rise to internal friction.

The damping is of the form

$$Q^{-1} = \frac{C_p - C_v}{C_v} \cdot \mathscr{R} \cdot \frac{f_0 f}{f_0^2 + f^2}, \qquad \ldots\ldots(72)$$

where $\mathscr{R}$ is that fraction of the total strain energy which is associated with fluctuations in dilation. The relaxation frequency is found by dimensional analysis to be $f_0 \sim D/L_c^2$, where D is the thermal diffusion constant and L_c is the mean diameter of the crystallites. $\mathscr{R}$ is evaluated by considering a small cube divided in half; each half is a crystallite oriented at random. The result comes out in terms of Poisson's ratio and the elastic moduli. Values of $\mathscr{R}$, and $(C_p - C_v)/C_v$, and the product of these two factors are given in table 2 for shear waves in cubic crystals. Zener's value (1938) for lead is in error and is corrected in the table.

Table 2. Effect of elastic anisotropy in cubic crystals. (After Zener, 1938)

	Metal						
	Pb	Ag	Cu	Au	Fe	Al	W
$\mathscr{R}$	0·065	0·031	0·031	0·014	0·022	0·0009	10^{-6}
$(C_p - C_v)/C_v$	0·067	0·040	0·028	0·038	0·016	0·046	0·006
Product	$4{\cdot}4\times10^{-3}$	$1{\cdot}2\times10^{-3}$	$8{\cdot}7\times10^{-4}$	$5{\cdot}3\times10^{-4}$	$3{\cdot}5\times10^{-4}$	4×10^{-5}	6×10^{-9}

The last line in the table gives an idea of the comparative damping in the various metals when the frequencies are chosen so that the ratio f/f_0 is the same in all cases. It is known experimentally from various preliminary measurements around 10 Mc/sec. that polycrystalline lead and copper are highly absorbing, while aluminium and tungsten are fairly good conductors of sound waves. It will be seen below that scattering losses also depend on the factor $\mathscr{R}$, so that it is not clear without further consideration whether scattering or thermoelastic internal friction is likely to be the dominant cause of absorption around 10 Mc/sec.

Extensive references are given by Seitz (1943) to experimental measurements in the 1–100 kc/sec. frequency range in support of Zener's theory.

(iii) *Scattering*

The fact that the strain in a vibrating polycrystalline solid varies incoherently from crystallite to crystallite means that energy is lost from a sound beam by scattering processes. The scattering arises principally because of the inhomogeneous shearing motion associated with both longitudinal and shear waves. In a cubic crystal a macroscopic pure dilation implies a uniform dilation of all crystallites, so that there should be no scattering for a pure dilation.

The scattering calculation has been carried out by Mason and McSkimin (1947) on lines similar to Rayleigh's theory (1894) of sound scattering in gases caused by inhomogeneities in the medium. The assumptions underlying the Rayleigh λ^{-4} law are that: (*a*) the wavelength must be long compared with the size of the inhomogeneity; (*b*) the local fractional variations in the character of the medium must be small; and (*c*) multiple scattering may be neglected.

For shear waves the Rayleigh absorption is given by

$$\alpha = (4\pi^3 L_c^3 \mathscr{R}/3\lambda^4), \qquad \ldots\ldots(73)$$

where L_c^3 is the mean volume of the crystallites and $\mathscr{R}$ is the factor introduced by Zener (1938) in the discussion of thermoelastic internal friction to denote that fraction of the total strain energy which is associated with fluctuations in

dilation. $\mathscr{R}$ may be considered as a measure of the *scattering power* of a material.

Reference to values of $\mathscr{R}$ given in table 2 shows, for example, that copper has a high scattering power, while aluminium and tungsten have low scattering powers. It is known experimentally that the absorptions in these metals have the same qualitative relation as the scattering powers. The only quantitative measurements available on specimens of known grain size are those of Mason and McSkimin (1947) on polycrystalline aluminium rods with average grain dimensions of 0·010 and 0·005 inch. The experimental results are in quite satisfactory agreement with equation (73) with regard to the dependence of the absorption on frequency and grain size, and the value of $\mathscr{R}$ derived from the attenuation coefficient for shear waves is in good agreement with the value calculated from the elastic constants as given in table 2 above. It appears, however, that there may be some difficulty in accounting for the absorption coefficient for longitudinal waves in the aluminium samples.

Mason and McSkimin find that for both shear and longitudinal waves the absorption over a wide range of frequencies can be represented as the sum of a term linear in f and a term in f^4:

$$\alpha = B_1 f + B_2 f^4. \qquad \ldots\ldots(74)$$

The B_1 term is ascribed to elastic hysteresis and the B_2 term to Rayleigh scattering. The term in B_1 does not seem to be sensitive to either grain size or the type of wave considered. In fused quartz and glasses in the 1 to 20 Mc/sec. range there does not seem to be any Rayleigh scattering present, and here all of the absorption is thought to be accounted for by elastic hysteresis.

The order of magnitude of the frequency at which $\lambda = L_c$ should be noted. Supposing a wave velocity of $5 \cdot 10^5$ cm/sec., for $L_c = 0{\cdot}1$, 0·01, and 0·001 cm. one has $f = 5$, 50, and 500 Mc/sec., respectively. In general the frequencies at which the thermoelastic effect is important are considerably lower than the frequencies at which scattering is important.

The Rayleigh λ^{-4} scattering law breaks down when the wavelength is of the same order as or is smaller than the grain size (Pekeris, 1947). In the limit of short wavelengths the propagation of sound probably becomes similar to a diffusion process.

(iv) *Plastic flow and elastic hysteresis*

Elastic hysteresis as used here is a poorly defined term understood to mean energy losses in solids not definitely tracked down to specific mechanisms. It includes such factors as plastic flow and the viscous slip of grain boundaries.

Wegel and Walther (1935) used a resonance method to determine the Q's of commercial materials at frequencies up to 500 kc/sec. They find in general that the Q's are roughly independent of frequency. It follows from equation (71) that the absorption coefficient must be approximately proportional to the first power of the frequency in the range studied; it will be remembered that Mason and McSkimin find a term linear in f which would be dominant at the lower frequencies. A term linear in f implies that the mechanical hysteresis loop is independent of frequency, so that the energy loss per cycle is constant.

It is difficult to see what type of mechanism is responsible for the term linear in f. It is not an ordinary viscosity, since viscous mechanisms give terms

quadratic in f. Relaxation mechanisms give terms quadratic in f at low frequencies and independent of f at high frequencies.

Read (1940) finds that the logarithmic decrement of *annealed single metal crystals* varies inversely with the frequency, and is of the order of 10^{-4} to 10^{-5} for frequencies of the order of 50 kc/sec. The absorption coefficient α is independent of the frequency, which suggests that the origin of the internal friction studied by Read is the same as that of macroscopic plastic flow (Houwink, 1937; Zener, 1946). The relaxation frequency for plastic flow in metals is of the order of 0·0001 cycles/sec., so that the working frequency satisfies $f \gg f_0$. In this limit the attenuation coefficient for relaxation processes is independent of the frequency, as is seen for example from equation (32). Plastic flow has been explained as being caused by the propagation of dislocations through the lattice. The observed dependence of the absorption on stress amplitude and temperature is in qualitative agreement with this explanation.

In polycrystals an additional cause of loss is provided by the viscous behaviour of grain boundaries (Kê, 1947; West, 1946).

(v) *Thermal conductivity*

The absorption of sound caused by thermal conductivity is given by equation (15), as for gases and liquids. It may be a significant factor in single metal crystals at frequencies above 1000 Mc/sec.; for example, in aluminium at 3000 Mc/sec. the calculated value of α caused by thermal conductivity alone is of the order of 100 nepers/cm., so that the pressure amplitude falls by $1/e$ in 0·01 cm.

(vi) *Structural and thermal relaxation*

Absorption caused by thermal relaxation on a molecular scale may possibly be one of the causes of the absorption observed in plastics (Mason, 1942) and rubber-like materials. It is possible that the study of absorption in crystals of simple organic compounds would produce information about the connection between relaxation frequencies in solids and liquids, but it seems that no work has been done in this field.

Tisza has suggested that it would be interesting to study relaxation absorption near λ-point transitions in solids. Lawson (1940) has studied the peak in absorption as a function of temperature in polycrystalline ammonium chloride near the transition at 242·8 °K., but not over a range of frequencies.

(vii) *Anharmonic coupling*

Energy losses caused by anharmonic coupling between the sound wave and the thermal lattice vibrations have been the subject of calculations by Landau and Rumer (1937), Akhieser (1939), Pomeranchuk (1941), and Gurevich (1945). The theoretical calculations apply only to high hypersonic frequencies and it is somewhat questionable whether losses of this character can be detected experimentally.

(viii) *Magnetic effects*

A good review of the experimental and theoretical material on internal friction in ferromagnetic materials has been given by Becker and Döring (1939)

The problem is of considerable importance in connection with losses in magnetostrictive transducers.

The dependence of the internal friction on magnetization is very striking. The internal friction in a nickel specimen (Siegel and Quimby, 1936) magnetized to saturation has values characteristic of non-magnetic materials and is not affected by heating above the Curie point. If the specimen is *demagnetized* in a macroscopic sense, the damping below the Curie point is increased by a factor of the order of 50; and when the Curie point is approached from below the damping rapidly decreses.

The energy losses in ferromagnetic materials are ascribed to magneto-mechanical hysteresis and also to *eddy currents* induced by *magnetostrictive changes in magnetization* caused by the stress associated with a sound wave.

(ix) *Recent work in solids*

Among the recent experimental studies of sonic and ultrasonic propagation in solids are the following:

Binary alloys.	Allan (1941).
Lucite.	Weber and Goeder (1942); Mason (1942).
Aluminium oxide ("Degussit")	Thiede (1941).
Effect on X-ray diffraction.	Weigle and Bleuler (1942); theory by Weigle (1941).
Rock salt near melting point.	Hunter and Siegel (1942).
Effect on magnetic behaviour of nickel.	Jetter (1941); Schmidt and Jetter (1942).
Delay line materials.	Arenberg (1945).
Propagation in wires.	Czerlinski (1942).
Elastic constants of alkali halide crystals.	Nurmi (1941); Huntington (1947); Galt (1947).
Rubberlike materials.	Nolle (1947).

The German material "Degussit" (Thiede, 1941) is noteworthy because of its unusually high sound velocity of 9600 m/sec.

The study by Hunter and Siegel (1942) of the variation of the shear modulus in rock salt in the immediate neighbourhood of the melting point is of theoretical interest in connection with lattice dynamics.

Smekal (1940) has discussed the connection between speed of fracture in solids and a possible cause of dispersion in the hypersonic region.

A comprehensive review of the elastic constants of anisotropic materials has been given by Hearmon (1946).

Galt (1947) has measured the elastic constants of KBr from room temperature to liquid helium temperature by an ultrasonic pulse method. He finds $c_{11}=4{\cdot}18\times10^{11}$ dyne/cm^2 at 4·5° K. and $c_{11}=3{\cdot}46\times10^{11}$ dyne/cm^2 at 300° K.; on the other hand, c_{44} is substantially constant over this range. He has also determined strain-optical constants in NaCl, KCl and KBr.

A great deal of work on propagation in solids has been done in connection with delay lines which are used to store information in the form of ultrasonic pulses

for times of the order of 10^{-4} sec. Much of this work is unpublished, but Arenberg's report (1945) is a useful source of information.

With regard to experimental methods, the resonance techniques employed below one megacycle have been reviewed by Siegel (1944). The measurement of absorption by the pulse method is discussed by Mason and McSkimin (1947).

(x) *Present status of measurements in solids*

The thermoelastic damping appears to be quite well understood as a result of work by Zener and his collaborators. The scattering theory accounts fairly well for the general features of the absorption in polycrystalline materials in the range above one megacycle. Some of the features of the damping in ferromagnetic materials are also understood. Other aspects of propagation in solids have not been studied in sufficient detail to make an intelligent comparison with what theories there are. It seems likely that the study of solids may be the most profitable field in ultrasonics research in the immediate future.

§5. EXPERIMENTAL METHODS

(i) *Pulse technique*

The outstanding recent experimental development in ultrasonics is the application of electronic pulsed-circuit techniques in the frequency range above one megacycle. The major tools of ultrasonics now include the optical diffraction method, the acoustic interferometer method, and the pulse method.

Pulsed-circuit ultrasonic methods have been used for some time in underwater echo-ranging equipment operating in the region of 10 kc/sec. The development of radar gave an impetus to work in the 1 to 30 Mc/sec. range, where ultrasonics was required for a number of purposes. These purposes included the use of ultrasonic delay lines to "store" radar signals for periods of time up to several milliseconds in connection with range measuring devices, moving target discriminators and test equipment; ultrasonics was used also in synthetic radar training equipment, where the appearance of terrain on radar displays was simulated with "small scale terrain" swept by an ultrasonic beam.

The use of the delay lines and training devices depends on the fact that the ratio of the velocity of electromagnetic radiation in air to the velocity of sound in a liquid or solid is of the order of 10^5, so that, for example, a 10-km. radar path is simulated in travel time by a 10-cm. sound path, and a radar wavelength of 10 cm. at 3000 Mc/sec. scales down to an acoustic wavelength of $\sim$0·01 cm. at 30 Mc/sec.

Because of the circumstances surrounding the introduction of ultrasonic pulse technique it is difficult to give adequate credit to those who participated in the development. Early workers include W. Shockley of the Bell Telephone Laboratories in the U.S.A., H. Grayson of the Telecommunications Research Establishment in Britain, and, apparently, F. Kruse of Telefunken in Germany. The M.I.T. Radiation Laboratory was also very active in the field.

The application of pulse technique to the study of the velocity and absorption of ultrasonic waves has been reported on by Arenberg (1945), Cefola, Droz, Frankel, Jones, Maslach, and Teeter (1946), Teeter (1946), Pellam and Galt (1946) and Rapuano (1947).

The equipment used by Pellam and Galt in their study of propagation in liquids at 15 Mc/sec. will be described here. The method is essentially to use the liquid sample as a "storage medium" for short sound pulses and to measure electrically the time delay and attenuation undergone by the sound pulse in traversing a known path within the liquid. Pulses of about one micro-second duration are generated by a quartz transducer, which also picks up the resultant echoes from a fixed plane reflector. Velocity measurements are made by determining the distance the transducer must be moved to delay the received echoes by a specified increment. Absorption measurements are made by determining the attenuation necessary to keep the receiver signal constant as the transducer is moved. The absorption coefficient was measured to an accuracy of about 5% and sound velocity to about 0·05%.

A block diagram of the equipment is shown in figure 2. The timing sequence

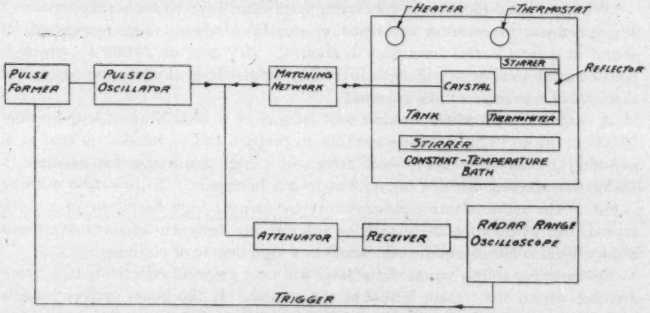


Figure 2. Block diagram of pulse equipment.

(Reproduced, by kind permission, from the 'Journal of Chemical Physics'.)

begins when a crystal-controlled circuit in the oscilloscope starts the oscilloscope sweep and simultaneously sends a trigger signal to the pulse generator. The pulse formed by the generator passes through an impedance matching network to the transducer, which produces a sound pulse in the medium under study. The sound pulse is returned to the transducer by a reflector, and a delayed electrical pulse is formed which passes back through the matching network. The transmitter is off at this time, and the main portion of the pulse travels through the attenuator to the receiver.

The output of the receiver is fed to an oscilloscope equipped with a special sweep indicating time delays; the start of the sweep is synchronized with the generation of the original pulse. The sweep is calibrated accurately with respect to time. The change of the signal position on the screen as the transducer mount is moved relative to the reflector is a direct measure of the corresponding time delay; simultaneously, electrical attenuation is inserted or removed by means of the calibrated attenuator so as to keep the receiver input-signal constant.

The crystal radiating surface is about 100 wavelengths across, so that the divergence of the beam is less than 1° and for most purposes may be neglected.

Errors in the measurement of absorption coefficient amounting to less than 1% are introduced by the fact that a finite band of frequencies must be used to form a pulse. The pulses employed by Pellam and Galt contained about 15 oscillations. For very high attenuations the returned echo is often obscured by the transmitted pulse; Rapuano (1947) has circumvented this difficulty by using an acoustic delay line as the reflector.

(ii) *Progress towards higher frequencies*

Advances in electronics connected with the development of microwave radar have made it possible to generate large amounts of electromagnetic power at frequencies up to 30000 Mc/sec., corresponding to 1-cm. electromagnetic wavelength. On the other hand, no ultrasonic measurements have been reported above 1000 Mc/sec. (Ringo, Fitzgerald and Hurdle 1947).

What are the difficulties in generating ultrasonic waves at higher frequencies? It is pertinent to consider the situation at 3000 Mc/sec. The wavelength of sound in quartz at this frequency is about 2×10^{-4} cm., or 20000 A., which is just 3 or 4 times the wavelength of visible light. It is clear in the first place that *optical tolerances* will be required.

A quartz crystal whose fundamental frequency is 3000 Mc/sec. will be only 10000 A. thick. This is much too thin to prepare and to handle, so that it is necessary to use a thicker crystal driven in a high harmonic, for example, a 30-Mc/sec. crystal driven near its hundredth harmonic. In principle nothing is lost at the same electric-field intensity by using a high harmonic of a thick crystal, but in practice there may be interference between adjacent overtones if the crystal is not of uniform thickness to a high degree of accuracy.

In receiving sound waves the crystal will only respond effectively to a beam arriving within the "main lobe" of the crystal. If the beam arrives outside of the main lobe, different portions of the crystal face will be out of phase and the electrical pulse will be greatly weakened. For a crystal 1 cm. in diameter the angular position of the first minimum in the diffraction pattern at 3000 Mc/sec. occurs at about 0·01 degrees from the normal. This means that the alignment of the reflector relative to the transducer is very critical.

A second fundamental difficulty is the high absorption which obtains in all materials at high frequencies. One would expect that at 3000 Mc/sec. the pressure would be down by $1/e$ in $\sim 10^{-5}$ cm. in gaseous helium at S.T.P. and in $\sim 10^{-3}$ cm. in liquid mercury (which is a relatively low absorbing material). It is difficult to predict what will be the absorption in crystalline quartz, as there is little in the way of theory or experimental data at lower frequencies from which to extrapolate. D'Eustachio (1946) has reported that the surface layers of quartz crystals are slightly disturbed to a depth of $\sim 5 \times 10^{-4}$ cm., which suggests that scattering losses may enter. It is clear that large acoustic losses may arise at these frequencies from even thin films of absorbing material, and special experimental techniques may be required.

Preliminary theoretical consideration has been given by Slater and Pellam (unpublished) to the question of transducer design at microwave frequencies, and it would appear that in principle at least the problem can be solved by mounting the crystal between two resonant cavities.

There is evidence from dielectric constant measurements on quartz that the polarization mechanism operative at low frequencies is also effective in the microwave range, so that one may have confidence in the inherent piezoelectric properties of quartz at microwave frequencies.

(iii) *Delay lines*

The types of ultrasonic delay lines which have been developed include the following:

(*a*) Water or mercury in a long tube.
(*b*) Fused quartz block utilizing multiple reflections over an intricate path.
(*c*) Shear waves in a long fused quartz rod.

The principal requirement for a delay line material is low attenuation at 10 to 30 Mc/sec.

(iv) *Recent papers on experimental methods*

Recent publications on experimental methods in ultrasonics include the following:

Acoustic interferometer theory.	Hardy (1943); Krashnooshkin (1944); Born (1943).
Interferometer design.	Telfair and Pielemeier (1942); J. L. Stewart (1946).
Optical diffraction method.	Gollmick (1940); Bommel (1943-45); Petersen (1940); Bhagavantam and Ramachandra Rao (1946 a, 47); Gregg (1941); Houstoun (1944).
Generators.	Smith and Stumpf (1946); Ehret and Hahnemann (1942); Epstein *et al.* (1947).
Transducer theory.	Cady (1945); Swann (1946); Benfield, Emslie and Huntington (1944); Goudet (1941).
Electrostrictive transducer.	Falkenhagen and Mees (1944).
Thermal transducer.	Fairbank, Fairbank and Lane (1947).
Comparison of optical and interferometer methods.	Garten (1942).
Rigidity modulus in solids.	Kornfeld (1944).
Velocity in solids.	Prasad (1945 b, 46).
Elastic constants of small samples.	Bhagavantam and Ramachandra Rao (1946 b).
Isolation of crystal from conducting liquids, etc.	MacGrath and Kurtz (1942); Thiede (1943).
Wave front determination.	Labaw (1945).
Pulse method.	Arenberg (1945); Cesola *et al.* (1946); Teeter (1946); Pellam and Galt (1946).
Ultrasonic lenses.	Ernst (1945); Hopwood (1946); Willard (1947 b).
General reference.	Bergmann (1942).

§6. APPLICATIONS OF ULTRASONICS

(i) *Testing of materials*

The major development in the application of ultrasonics to fields other than physics is the use of the pulse technique in the testing of materials (Firestone, 1945; Simon, 1945). The ultrasonic pulse method promises to become one of the most important means of flaw detection, and it is already being used on an industrial scale.

Russell and Pellett (1946) have given an account of the use of the Sperry "supersonic reflectoscope" in the inspection of steel billets in a steel mill. The usual method of inspection is by means of fracture and macro-etch tests, both of which involve considerable labour and loss of material. The pulse technique is a useful supplementary method for the location of internal defects with internal reflection surfaces. Such defects include shrinkage cavities, internal cracks, and excessive inclusions.

The method of ultrasonic flaw detection consists in sending a pulse of sound at a frequency usually between 1 and 5 Mc/sec. through the specimen. The pulse may be a shear, longitudinal, or Rayleigh wave (Firestone and Frederick, 1946). If a material defect is present in the path of the sound pulse some of the energy of the pulse will be scattered or reflected back to the transducer. The received signal is displayed on a calibrated oscilloscope. The pulse will also be reflected back from the other end of the specimen.

It is usually necessary to grind the surface of the specimen to insure good acoustic contact. In the case of longitudinal waves a film of oil is used between the crystal and the surface of the specimen.

(ii) *Recent applications of ultrasonics*

Some recent applications of ultrasonics to fields other than physics are listed below. The Report by Richardson (1944-45) may be consulted for references to metallurgical applications.

Material testing.	Firestone (1945); Simon (1945); Delano (1946); Sokolov (1941); Trent (1945); Russell and Pellett (1946); Firestone and Frederick (1946).
Gas analysis.	Vengerov (1946).
Pressure gauges.	Fox, Herzfeld and Rock (1946).
Ceramics research.	Bole and Loomis (1943).
Colloid research.	Sollner (1944); Rutgers (1946).
Biological research.	Schubotz (1941); Lynn, Zwemer and Chick (1942); Lynn, Zwemer, Chick and Miller (1942); Klöpf (1941).
Metal coating.	Townsend (1946).

§7. SOUND PROPAGATION IN LIQUID HELIUM

It has recently been discovered that wave motion can be propagated through helium II, the low temperature phase of liquid helium, with two distinct velocities. One motion corresponds to the ordinary pressure wave and has a velocity of the order of 240 m/sec. The other motion, which is called the "second sound",

represents the propagation of a temperature wave and has a velocity strongly dependent on temperature with a maximum of 20 m/sec. around 1·6° K.

The pressure wave or first sound may be transmitted and received by an electro-acoustic transducer in the usual way. The temperature wave or second sound is generated by means of an alternating current heater and is received by a thermal detector such as a phosphor-bronze resistance thermometer.

The second sound was predicted from the theory of quantum liquids independently by Tisza (1938-40) and Landau (1941-4). The existence of the second sound was demonstrated experimentally by Peshkov (1944, 46), acting on Landau's suggestion. The prior prediction of this effect is a great achievement of the theory.

Helium II is that phase of liquid helium which exists below about 2·2° K. It is believed that helium II, under its own vapour pressure, remains liquid down to absolute zero, although it solidifies when a pressure of about 25 atmospheres is applied. The unusual properties of helium II arise from the fact that it is a "quantum liquid"; that is, a liquid in which the quantum mechanical corrections to classical statistical mechanics become of major significance. This means that the zero point motion of the liquid is important.

The theories of Landau and Tisza are based on different molecular models, but lead to similar, though not identical, macroscopic conclusions. Both theories arrive at the result that in some respects helium II may be considered as a mixture of two different liquids, of which one behaves as a *normal* fluid and the other as a *superfluid*. The superfluid is characterized by zero viscosity and there is therefore no friction between the two liquids flowing with respect to each other. The entropy of the superfluid component may be considered to be zero.

In sound waves of the first type the mixture of liquids vibrates as a whole in each volume element, and these waves correspond to ordinary sound waves in ordinary liquids. In a wave of the second type the superfluid and normal components of the liquid vibrate opposite to each other in such a manner that in the first approximation the centre of gravity of each volume element remains stationary and there is no total flow of matter. In waves of this type the pressure amplitude is small and the temperature amplitude is large, whereas in the first sound the opposite is true.

The idea is sometimes advanced (Richardson, 1946) that the second sound can be understood as a case of the so-called heat wave which occurs in all media when an alternating temperature is imposed on one boundary. This is a mistaken conception. The wave vector for a heat wave is of the same form as equation (55) for a shear wave, provided that the kinematic viscosity η/ρ is replaced by the thermal diffusion constant D. The wave velocity given by equation (57) is seen to be proportional to $\omega^{\frac{1}{2}}$, and the attenuation is such that the amplitude diminishes by $1/e$ in the distance $\lambda/2\pi$. Both of these statements are in contradiction to the observed absence of dispersion and the apparently small absorption.

Lifshitz (1944) has shown that the coupling at an ordinary acoustic transducer between first and second sound is very small: the ratio of the radiated intensity of second sound to first sound is of the order of 10^{-8}. On the other hand, the radiation from a surface whose temperature fluctuates periodically in time is favourable for the second sound by a factor of the order of 10^4 in intensity.

Lane, Fairbank, Schultz and Fairbank (1946–7) have measured the velocity of second sound using a method suggested by Onsager. The alternating temperature caused by second sound incident on a liquid-vapour interface causes periodic evaporation which in turn gives rise to normal sound in the vapour. Onsager estimates that the transmission across the liquid surface would be of the order of 70%. The normal sound is detected by a microphone. The second sound was generated in the liquid by a flat resistance heater of fine constantan wire fed with 1000-cycle A.C. The second sound velocity at 1·74° K. was 1918 cm/sec., confirming Peshkov's earlier result.

The relation between the Landau and Tisza theories has been recently discussed by Tisza (1947). It appears that the macroscopic properties of helium II can be derived from rather general quasi-thermodynamic considerations. An expression for the velocity of propagation of second sound is found in terms of a single empirical function, the entropy, without further adjustable constants. The earlier theory of Tisza based on London's Bose-Einstein gas theory appears as a special molecular model, which is on the whole compatible with the thermodynamic theory. Landau's theory in its present form has features which conflict with the latter.

Landau predicts that as the temperature approaches zero, the second sound velocity will approach $(\frac{1}{3})^{\frac{1}{2}}$ times the first sound velocity, while the thermodynamic theory confirms Tisza's earlier result that the second sound velocity approaches zero as T approaches zero. Landau's velocity values are consistently higher than those computed by Tisza. The experiments tend to support Tisza's theory.

It appears, however, that the hydrodynamic approach of Landau can be so modified as to be free from these objections. The agreement between the conclusions drawn from the Bose-Einstein molecular model and the modified hydrodynamic model seems to be so close that a decision on the basis of present experiments is hardly possible.

Measurements have been made by Pellam and Squire (1947) on the velocity and attenuation of first sound in liquid helium at a frequency of 15 Mc/sec. as a function of temperature from 1·57° K. to 4·5° K. The measurements exhibit three important features: (*a*) in the upper temperature range of He I the measured attenuation agrees very well with classical theory; (*b*) at the λ-point the attenuation rises abruptly, presumably to infinity, indicating complete absorption of the ultrasonic energy; (*c*) just below the λ-point the attenuation has its minimum value, and with lowering temperature the attenuation increases.

APPENDIX I

Notations and symbols

a_H = Bohr radius = $0{\cdot}53 \times 10^{-8}$ cm.
C = effective elastic modulus.
C_p, C_v = molar heat capacities at constant pressure and constant volume.
c_p, c_v = heat capacities per gram at constant pressure and constant volume.
c = velocity of sound = ω/k_1.
c' = molecular velocity.

C_i, C_e = molar heat capacities of internal and external degrees of freedom.
D = thermal diffusion constant $= K/\rho c_p$
e = strain.
E = Young's modulus.
F = force between molecules.
f = frequency.
f_0 = relaxation frequency.
h = Planck's constant $/2\pi = 1{\cdot}05 \times 10^{-27}$ erg sec.
$j = (-1)^{\frac{1}{2}}$.
K = coefficient of thermal conductivity.
k = Boltzmann's constant.
$k = k_1 - jk_2 = (2\pi/\lambda - j\alpha)$ = complex wave vector.
k_2 = pressure attenuation coefficient $= \alpha$.
L = intermolecular separation.
L_f = free length (see figure 1).
L_c = grain diameter.
m = mass of a molecule.
M = molecular weight.
N = number of molecules.
n = number of moles.
p = total pressure.
P = excess pressure caused by sound wave.
Q = thermal energy per mole.
Q = "Q" of system $= 2\pi$ (energy stored/energy dissipated per cycle).
R = gas constant.
$\mathscr{R}$ = Zener's constant (see table 2).
s = condensation $[\rho = \rho_0(1+s])$.
S = entropy.
S_{xy} = shear stress.
t = time.
T = Kelvin temperature.
U = internal energy per mole.
u = particle velocity.
v = relative molecular velocity.
V = volume.
V_a = available volume.
V_0 = minimum possible volume.
$W = MK/\rho\omega$ = parameter in equation (10).
W = molecular interaction energy.
α = absorption coefficient in nepers/cm. $[P = P_0 e^{-\alpha x}]$.
β = compressibility $= \dfrac{-1}{V}\left(\dfrac{\partial V}{\partial p}\right)_T$.
γ = ratio of heat capacities $= c_p/c_v$.
η = shear viscosity coefficient.
θ' = packing fraction.
θ = relative excess temperature $[T = T_0(1+\theta)]$.
κ = bulk modulus.
λ = wavelength of sound.
Λ = molecular mean free path.
μ = shear modulus.
Π = probability of energy exchange.
ρ = density.
ρ_0 = density corresponding to closest possible packing.
τ = relaxation time constant.
χ = compressional viscosity coefficient.
ω = angular frequency $= 2\pi f$.

APPENDIX II

Deformation and viscosity in plane waves

The effective elastic constant for the type of deformation occurring in a plane wave in an *infinite isotropic solid* will be derived here in a manner which is briefer and more physical than the usual rather formal derivation involving the general elasticity equations. The treatment will also include viscosity effects in liquids and gases.

Consider a unit cube on which a stress X_x acts to produce the relative length changes $e_x = e_x^0$, $e_y = e_z = 0$ in the x, y, z directions, respectively. This corre-

sponds to the character of the deformation in a plane wave whose front is normal to the x direction. The problem is to find the effective elastic modulus C for this deformation, where $X_x = Ce_x$. It is desired to express C in terms of the *bulk modulus* κ and the *shear modulus* μ, these being the two most revealing choices for the two elastic constants of an isotropic solid.

The final state of strain $e_x = e_x^0$, $e_y = e_z = 0$ may be considered as reached in three successive stages:

(i) Uniform compression to $e_x = e_y = e_z = e_x^0/3$.
(ii) An x-push and a y-pull at constant volume to $e_x = 2e_x^0/3$; $e_y = 0$; $e_z = e_x^0/3$.
(iii) An x-push and a z-pull at constant volume to $e_x = e_x^0$; $e_y = e_z = 0$.

Stage 1. The uniform compression requires a stress $X_x = Y_y = Z_z = \kappa e_x^0/3$; the stresses Y_y and Z_z are of no concern as the net work done by the y and z forces in the entire 3-stage process must be zero.

Stage 2. In this deformation the x and y directions are given incremental strains of equal but opposite amounts $e_x^0/3$. Such a deformation is equivalent to a shear of angle $2e_x^0/3$, as may be seen by considering the stresses on planes making angles of $\pm 45°$ with the x axis (Timoshenko, 1934). The stress required is $X_x = 2\mu e_x^0/3 = -Y_y$.

Stage 3. The x and z directions are strained by equal but opposite increments $e_x^0/3$, requiring a stress $X_x = 2\mu e_x^0/3 = -Z_z$.

The total stress required to produce the resultant deformation $e_x = e_x^0$; $e_y = e_z = 0$ is the sum of the individual stresses, so that

$$X_x = (\kappa + \tfrac{4}{3}\mu)e_x^0,$$

giving an effective elastic modulus

$$C = \kappa + \tfrac{4}{3}\mu$$

for the plane wave type of deformation. The wave equation is

$$(\kappa + \tfrac{4}{3}\mu)(\partial^2 u/\partial x^2) = \rho(\partial^2 u/\partial t^2),$$

which gives equation (65) for the velocity.

Viscosity coefficients connect stress and time rate of change of strain, whereas elastic moduli connect stress and strain. By strict analogy one writes $\chi + \frac{4}{3}\eta$ for the effective viscosity coefficient in a plane wave, where χ is the *compressional viscosity coefficient* and η is the ordinary *shear viscosity coefficient*. The force equation on an element of volume is

$$\rho(\partial u/\partial t) - (\chi + \tfrac{4}{3}\eta)(\partial^2 u/\partial x^2) = -(\partial p/\partial x),$$

which gives an equation identical with equation (20) when the Stoke's assumption $\chi = 0$ is made.

REFERENCES

AKHIESER, A., 1939. *J. Phys. U.S.S.R.*, **1**, 277.
ALEXANDER, E. A. and LAMBERT, J. D., 1942. *Proc. Roy. Soc.*, A, **179**, 499.
ALLAN, G. E., 1941. *Phil. Mag.*, **32**, 165.
ALLEMAN, R. S., 1939. *Phys. Rev.*, **55**, 87.
ALLEMAN, R. S., 1941. *J. Acoust. Soc. Amer.*, **13**, 23.
ANSELM, A., 1945. *J. Exp. Theor. Phys. U.S.S.R.*, **15**, 751.
ARENBERG, D. L., 1945. *M.I.T. Radiation Laboratory Report*, **932**.
BACCAREDDA, M., 1946. *Ric. Sci.*, **16**, 662.

BACCAREDDA, M. and GIACOMINI, A., 1945. *Ric. Sci.*, **15**, 3.
BACCAREDDA, M. and GIACOMINI, A., 1946. *Ric. Sci.*, **16**, 611.
BÄR, R., 1937. *Helv. phys. Acta*, **10**, 311.
BARONE and GIACOMINI, A., 1945. *Ric. Sci.*, **15**, 2.
BATSCHINSKI, A. J., 1913. *Z. phys. Chem.*, **84**, 643.
BAZULIN, P. A., 1941. *C.R. Acad. Sci., U.R.S.S.*, **31**, 113.
BECKER, R. and DÖRING, W., 1939. *Ferromagnetismus* (Berlin: Verlag von Julius Springer; reprinted Ann Arbor: Edwards Bros.).
BENFIELD, A. E., EMSLIE, A. G. and HUNTINGTON, H. B., 1944. *M.I.T. Radiation Laboratory Report*, **792**.
BERGMANN, L., 1942. *Der Ultraschall und seine Anwendung in Wissenschaft und Technik* (Berlin: V.D.I.-Verlag; reprinted Ann Arbor: Edwards Bros.), 3rd edition.
BEUTTENMÜLLER, E., 1942. *Dissertation*, Technische Hochschule, Stuttgart.
BEYER, R. T. and SMITH, M. C., 1946. *J. Acoust. Soc. Amer.*, **18**, 424.
BHAGAVANTAM, S. and RAMACHANDRA RAO, B., 1946 a. *Nature, Lond.*, **158**, 484.
BHAGAVANTAM, S. and RAMACHANDRA RAO, B., 1946 b. *Proc. Indian Acad. Sci., Bangalore*, **A23**, 254.
BHAGAVANTAM, S. and RAMACHANDRA RAO, B., 1947. *Nature, Lond.*, **159**, 267.
BOLE, G. A. and LOOMIS, G. A., 1943. *J. Appl. Phys.*, **14**, 443.
BÖMMEL, H., 1943. *Helv. phys. Acta*, **16**, 423.
BÖMMEL, H., 1945. *Helv. phys. Acta*, **18**, 3.
BORN, H., 1943. *Z. Phys.*, **120**, 383.
BRILLOUIN, L., 1922. *Ann. Phys., Paris*, [9] **17**, 88.
BRILLOUIN, L., 1933. *Actualites Scientifiques et Industrielles*, **59**.
BRILLOUIN, L., 1936. *J. Phys. Radium* [7], **7**, 153.
CADY, W. G., 1945. *Wesleyan University Report to Div. 14 NDRC.*
CEFOLA, M., DROZ, M. E., FRANKEL, S., JONES, E. M., MASLACH, G. and TEETER, C. E., Jr., 1946. *M.I.T. Radiation Laboratory Report*, **963**.
CONDON, E. U., 1933. *Amer. Phys. Teacher*, **1**, 18.
COREY, V. B., 1943. *Phys. Rev.*, **64**, 350.
CURTIS, R. W., 1934. *Phys. Rev.*, **46**, 811.
CZERLINSKY, E., 1942. *Akust. Z.*, **7**, 12.
DEBYE, P., 1939. *Z. Elektrochem.*, **45**, 174.
DELANO, R. B., Jr., 1946. *Electronics*, **19**, 132.
DERENZINI, T. and GIACOMINI, A., 1942 a. *Ric. Sci.*, **13**, 27.
DERENZINI, T. and GIACOMINI, A., 1942 b. *Ric. Sci.*, **13**, 542.
D'EUSTACHIO, D., 1946. *Phys. Rev.*, **70**, 522.
DUTTA, A. K. and GHOSH, B. B., 1944. *Indian J. Phys.*, **17**, 19.
ECKHART, C., 1947. *Phys. Rev.*, **71**, 277 A.
EHRET, L. and HAHNEMANN, H., 1942. *Z. tech. Phys.*, **23**, 245.
EPSTEIN, L. F., ANDERSON, W. M. A. and HARDEN, L. R., 1947. *J. Acoust. Soc. Amer.*, **19**, 248.
ERNST, P. F., 1945. *J. Sci. Instrum.*, **22**, 238.
EYRING, H. and HIRSCHFELDER, J. O., 1937. *J. Phys. Chem.*, **41**, 249.
FAIRBANK, H. A., FAIRBANK, W. M. and LANE, C. T., 1947. *J. Acoust. Soc. Amer.*, **19**, 475.
FALKENHAGEN, H. and MEES, R., 1944. *Reichsberichte für Physik* (restricted supplement to the *Physikalische Zeitschrift*), **1**, 42.
FERRY, J. D., 1941. *Rev. Sci. Instrum.*, **12**, 79.
FERRY, J. D., 1942 a. *J. Amer. Chem. Soc.*, **64**, 1323.
FERRY, J. D., 1942 b. *J. Amer. Chem. Soc.*, **64**, 1330.
FIRESTONE, F. A., 1945. *Metal Progress*, **48**, 505.
FIRESTONE, F. A. and FREDERICK, J. R., 1946. *J. Acoust. Soc. Amer.*, **18**, 200.
FOX, F. E., HERZFELD, K. F. and ROCK, G. D., 1946. *Phys. Rev.*, **70**, 329.
FOX, F. E. and ROCK, G. D., 1946. *Phys. Rev.*, **70**, 68.
FRANK, H. S., 1945. *J. Chem. Phys.*, **13**, 478.
FREI, H. and SCHIFFER, M., 1947. *Phys. Rev.*, **71**, 555.
FRENKEL, J., 1946. *Kinetic Theory of Liquids* (Oxford: Clarendon Press).
FRICKE, E. F., 1940. *J. Acoust. Soc. Amer.*, **12**, 245.

GALT, J. K., 1947. *Phys. Rev.*, **72** (in the press).
GARTEN, W., 1942. *Phys. Rev.*, **61**, 391 A.
GIACOMINI, A., 1942. *Acta d. Pont. Acc. delle Scienze* **6**, 87.
GINSBURG, V. L., 1942. *C.R. Acad. Sci., U.R.S.S.*, **36**, 8.
GOLLMICK, H. J., 1940. *Dissertation*, Berlin.
GORODETZKY, A., 1940. *J. Exp. Theor. Phys., U.S.S.R.*, **10**, 694.
GOUDET, G., 1941. *C.R. Acad. Sci., Paris*, **213**, 117.
GREGG, E. C., 1941. *Rev. Sci. Instrum.*, **12**, 149.
GROENEWOLD, H. J., 1939. *Physica*, **6**, 303.
GROSS, E., 1930. *Z. Phys.*, **63**, 685.
GUREVICH, L., 1945. *J. Phys. U.S.S.R.*, **9**, 383.
HALL, L., 1947. *Phys. Rev.*, **71**, 318 (L).
HARDY, H. C., 1943. *J. Acoust. Soc. Amer.*, **15**, 91.
HEARMON, R. F. S., 1946. *Rev. Mod. Phys.*, **18**, 409.
HERZFELD, K. F., 1935. *Ann. Phys., Lpz.*, **23**, 465.
HERZFELD, K. F., 1938. *Phys. Rev.*, **55**, 899.
HERZFELD, K. F., 1941. *J. Acoust. Soc. Amer.*, **13**, 33.
HERZFELD, K. F. and RICE, F. O., 1928. *Phys. Rev.*, **31**, 691.
HIEDEMANN, E., 1939. *Grundlagen und Ergebnisse der Ultraschallforschung* (Berlin : Walter de Gruyter).
HIRSCHFELDER, J. O., STEVENSON, D. P. and EYRING, H., 1937. *J. Chem. Phys.*, **5**, 896.
HOPWOOD, F. L., 1946. *J. Sci. Instrum.*, **23**, 63.
HOUSTOUN, R. A., 1944. *Phil. Mag.*, **35**, 192.
HOUWINK, R., 1937. *Elasticity, Plasticity, and Structure of Matter* (Cambridge : University Press).
HSU, E. T-Y., 1945. *J. Acoust. Soc. Amer.*, **17**, 127.
HUNTER, J. L., 1941. *J. Acoust. Soc. Amer.*, **13**, 36.
HUNTER, L. P. and SIEGEL, S., 1942. *Phys. Rev.*, **61**, 84.
HUNTINGTON, H. B., 1947. *Phys. Rev.*, **71**, 472 A.
ISAKOVICH, M. A., 1939. *C.R. Acad. Sci., U.R.S.S.*, **23**, 783.
JATKAR, S. K. M. and LAKSHIMINARAYANAN, D., 1946. *J. Indian Inst. Sci.*, **28** A, 1.
JEANS, J. H., 1925. *Dynamical Theory of Gases* (Cambridge : University Press), 4th ed.
JETTER, V., 1941. *Dissertation*, Stuttgart.
JONES, P. L. F. and GALE, A. J., 1946. *Nature, Lond.*, **157**, 341.
KALLMANN, H. and LONDON, F., 1929. *Z. phys. Chem.*, **B2**, 207.
KANTROWITZ, A., 1946. *J. Chem. Phys.*, **14**, 150.
KÊ, T.-S., 1947. *Phys. Rev.*, **71**, 533.
KELLER, H., 1940. *Phys. Z.*, **41**, 386.
KINCAID, J. F. and EYRING, H., 1937. *J. Chem. Phys.*, **5**, 587.
KINCAID, J. F. and EYRING, H., 1938. *J. Chem. Phys.*, **6**, 620.
KIRCHOFF, G., 1868. *Pogg. Ann.*, **134**, 177.
KITTEL, C., 1946. *J. Chem. Phys.*, **14**, 614.
KLOPF, A., 1941. *Dissertation*, Munich.
KNESER, H. O., 1938. *Ann. Phys., Lpz.*, **32**, 277.
KNESER, H. O., 1941. *Ann. Phys., Lpz.*, **39**, 261.
KNUDSEN, V. O., and FRICKE, E., 1940. *J. Acoust. Soc. Amer.*, **12**, 255.
KOHLER, M., 1941. *Ann. Phys., Lpz.*, **39**, 209.
KORNFELD, M., 1944. *J. Phys. U.S.S.R.*, **8**, 4.
KRASHNOOSHKIN, P. E., 1944. *Phys. Rev.*, **65**, 190.
LABAW, L. W., 1945. *J. Acoust. Soc. Amer.*, **17**, 19.
LABAW, L. W. and WILLIAMS, A. O., Jr., 1947. *J. Acoust. Soc. Amer.*, **19**, 30.
LAGEMANN, R. T., 1944. *J. Chem. Phys.*, **12**, 464.
LAGEMANN, R. T. and CORRY, J. E., 1942. *J. Chem. Phys.*, **10**, 759.
LAGEMANN, R. T. and DUNBAR, W. S., 1945. *J. Phys. Chem.*, **49**, 428.
LANDAU, L., 1941. *J. Phys., U.S.S.R.*, **5**, 71.
LANDAU, L., 1944. *J. Phys., U.S.S.R.*, **8**, 1.
LANDAU, L. and RUMER, G., 1937. *Phys. Z. Sowjet*, **11**, 18.
LANDAU, L. and TELLER, E., 1936. *Phys. Z. Sowjet*, **10**, 34.
LANE, C. T., FAIRBANK, H. A. and FAIRBANK, W. M., 1947. *Phys. Rev.*, **71**, 600.

LANE, C. T., FAIRBANK, H. A., SCHULTZ, H. and FAIRBANK, W, M., 1946. *Phys. Rev.*, **70**, 431 L.
LAWSON, A. W., 1940. *Phys. Rev.*, **57**, 417.
LIEPMANN, H. W., 1936. *Helv. phys. Acta*, **9**, 507.
LIFSHITZ, E., 1944. *J. Phys. U.S.S.R.*, **8**, 110.
LUCAS, R., 1937. *J. Phys. Radium* [7], **8**, 410.
LUCAS, R. and BICQUARD, P., 1937. *Trans. Faraday Soc.*, **33**, 130.
LYNN, J. G., ZWEMER, R. L. and CHICK, A. J., 1942. *Science*, **96**, 119.
LYNN, J. G., ZWEMER, R. L., CHICK, A. J. and MILLER, A. E., 1942. *J. Gen. Physiol.*, **26**, 179.
MCGRATH, J. W. and KURTZ, A. R., 1942. *Rev. Sci. Instrum.*, **13**, 128.
MANDELSTAM, L. and LEONTOVIC, M., 1936. *C.R. Acad. Sci., U.R.S.S.*, **3**, 111.
MANDELSTAM, L. and LEONTOVIC, M., 1937. *J. Exp. Theor. Phys., U.S.S.R.*, **7**, 438.
MARK, H., 1945. *J. Acoust. Soc. Amer.*, **16**, 183.
MASON, W. P., 1942. *N.D.R.C. Report* 6.1—sr. 346–628 (Dec. 16).
MASON, W. P., 1942. *Electromechanical Transducers and Wave Filters* (New York : D. Van Nostrand Co., Inc.), p. 247.
MASON, W. P., 1946. *Phys. Rev.*, **70**, 110A.
MASON, W. P., 1947. *Trans. Amer. Soc. Mech. Engrs.*, **69**, 359.
MASON, W. P. and MCSKIMIN, H. J., 1947. *J. Acoust. Soc. Amer.*, **19**, 464.
MATTA, K. and MOKHTAR, M., 1944. *J. Acoust. Soc. Amer.*, **16**, 120.
MAXWELL, J. C., 1867. *Phil. Trans.*, **157**, 49.
MIKHAILOV, I. G. and GOUREVITCH, S. B., 1946. *C.R. Acad. Sci., U.R.S.S.*, **52**, 673.
MOKHTAR, M. and RICHARDSON, E. G., 1945. *Proc. Roy. Soc.*, A, **184**, 117.
NATANSON, L., 1901. *Z. phys. Chem.*, **38**, 690.
NOLLE, A. W., 1947. *J. Acoust. Soc. Amer.*, **19**, 194.
NOURY, J., 1946. *C.R. Acad. Sci., Paris*, **223**, 377.
NURMI, U., 1941. *Soc. Sci. fenn. Comm. phys. math.*, **11**, 1.
OSIDA, I., 1941. *Proc. Phys.-Math. Soc., Japan*. **23**, 18.
OVERBECK, C. J. and KENDALL, H. C., 1941. *J. Acoust. Soc. Amer.*, **13**, 26.
PARSHAD, R., 1947. *J. Chem. Phys.*, **15**, 418.
PÄSLER, M., 1944. *Z. Phys.*, **122**, 357.
PEKERIS, C. L., 1947 a. *Phys. Rev.*, **71**, 268 L.
PEKERIS, C. L., 1947 b. *Phys. Rev.*, **71**, 457 L.
PELLAM, J. R. and GALT, J. K., 1946. *J. Chem. Phys.*, **14**, 608.
PELLAM, J. R. and SQUIRE, C. F., 1947. *Phys. Rev.*, **71**, 477 A.
PESHKOV, V., 1944. *J. Phys. U.S.S.R.*, **8**, 381.
PESHKOV, V., 1946. *J. Phys. U.S.S.R.*, **10**, 389.
PETERSEN, O., 1940. *Phys. Z.*, **41**, 29.
PFRIEM, H., 1940 a. *Akust. Z.*, **5**, 202.
PFRIEM, H., 1940 b. *Akust. Z.*, **5**, 103.
PFRIEM, H., 1941. *Akust. Z.*, **6**, 109.
PIELEMEIER, H., 1943. *J. Acoust. Soc. Amer.*, **15**, 22.
PIELEMEIER, H., 1945. *J. Acoust. Soc. Amer.*, **17**, 24.
PIELEMEIER, H. and BYERS, W. H., 1943. *J. Acoust. Soc. Amer.*, **15**, 17.
POISSON, S. D., 1831. *J. de l'Ecole Polytechn*, **20**, 139.
POMERANCHUK J., 1941, *J. Phys. U.S.S.R.*, **4**, 529.
PRASAD, R., 1941. *Indian J. Phys.*, **15**, 323.
PRASAD, R., 1942 a. *Indian J. Phys.*, **16**, 1.
PRASAD, R., 1942 b. *Indian J. Phys.*, **16**, 370.
PRASAD, R., 1945 a. *Indian J. Phys.*, **19**, 47.
PRASAD, R., 1945 b. *Nature, Lond.*, **156**, 637.
PRASAD, R., 1946. *Nature, Lond.*, **158**, 789.
PRIMAKOFF, H., 1942. *J. Acoust. Soc. Amer.*, **14**, 14.
PUMPER, E. J., 1939. *J. Phys., U.S.S.R.*, **1**, 411.
QUIGLEY, T. H., 1945. *Phys. Rev.*, **67**, 298.
QUINN, J., 1946. *J. Acoust. Soc. Amer.*, **18**, 185.
RAMA RAO, M., 1946. *J. Chem. Phys.*, **14**, 699.

RAMAN, C. V. and VENKATESWARAN, C. S., 1939. *Nature, Lond.*, **143,** 798.
RAMM, W., 1936. *Phys. Z.*, **35,** 756.
RANK, D. H., 1946. *J. Opt. Soc. Amer.*, **35,** 256.
RAPUANO, R. A., 1947. *Phys. Rev.*, **72,** 78.
RAYLEIGH, Lord, 1894. *Theory of Sound*, 2nd edition, Vol. II, p. 149 (London: Macmillan).
READ, T. A., 1940. *Phys. Rev.*, **58,** 371.
REIGER, R., 1910. *Ann. Phys., Lpz.*, **31,** 51.
RENDALL, G. R., 1943. *Proc. Indian Acad. Sci., Bangalore*, **A16,** 369.
RHODES, J. E., Jr., 1946. *Phys. Rev.*, **70,** 932.
RICE, O. K., 1942. *J. Chem. Phys.*, **10,** 653.
RICE, O. K., 1944. *J. Chem. Phys.*, **12,** 1.
RICE, O. K., 1946. *J. Chem. Phys.*, **14,** 335.
RICHARDS, W. T., 1939. *Rev. Mod. Phys.*, **11,** 36.
RICHARDSON, E. G., 1944–45. *Rep. Progr. Phys.*, **10,** 120.
RICHARDSON, E. G., 1946. *Nature, Lond.*, **158,** 296.
RIECKMANN, P., 1939. *Phys. Z.*, **40,** 582.
RINGO, G. R., FITZGERALD, J. W., and HURDLE, B. G., 1947. *Phys. Rev.*, **72,** 87.
ROY, A. S. and ROSE, M. E., 1935. *Proc. Roy. Soc.*, A, **149,** 511.
RÜFER, W., 1942. *Ann. Phys., Lpz.*, **41,** 301.
RUSSELL, J. V. and PELLETT, H. E., 1946. *Iron Age*, **157,** 38.
RUTGERS, A. J., 1946. *Nature, Lond.*, **157,** 74.
SAKADI, Z., 1941. *Proc. Phys.-Math. Soc., Japan*, **23,** 208.
SANEYOSI, Z., 1941. *Electrotech. J., Tokyo*, **5,** 49.
SCHILLING, H. K., GIVENS, M. P., NYBORG, W. L., PIELEMEIER, W. A. and THORPE, H. A., 1947. *J. Acoust. Soc. Amer.*, **19,** 222.
SCHMID, G., 1940. *Z. phys. Chem.*, **A186,** 113.
SCHMIDT, E. and JETTER, U., 1942 a. *Z. Elecktrochem.*, **48,** 227.
SCHMIDT, E. and JETTER, U., 1942 b. *Z. Elektrochem.*, **48,** 513.
SCHUBOTZ, F., 1941. *Dissertation*, Berlin.
SEIDL, 1940. *Z. Phys.*, **116,** 359.
SEITZ, F., 1943. *Physics of Metals* (New York: McGraw-Hill Book Co.).
SIEGEL, S., 1944. *J. Acoust. Soc. Amer.*, **16,** 26.
SIEGEL, S. and QUIMBY, S. L., 1936. *Phys. Rev.*, **49,** 663.
SIMON, E. N., 1945. *Metal Progress*, **48,** 513.
SINGH, B. K., 1945. *Nature, Lond.*, **156,** 569.
SIVIAN, L. J., 1947. *J. Acoust. Soc. Amer.*, **19,** 254.
SMEKAL, A., 1940. *Phys. Z.*, **41,** 475.
SMITH, F. W., Jr. and STUMPF, P. K., 1946. *Electronics*, **19,** 116.
SOKOLOV, 1941. *J. Tech. Phys., U.S.S.R.*, **11,** 160.
SOLLNER, K., 1944. *Chem. Rev.*, **34,** 371.
STEWART, E. S., 1946. *Phys. Rev.*, **69,** 632.
STEWART, E. S., STEWART, J. L. and HUBBARD, J. C., 1945. *Phys. Rev.*, **68,** 231 L.
STEWART, J. L., 1946. *Rev. Sci. Instrum.*, **17,** 59.
STOKES, G. G., 1841. *Proc. Camb. Phil. Soc.*, **8,** 287.
SUNANDA BAI, K., 1942. *Proc. Indian Acad. Sci., Bangalore*, **A15,** 349.
SURYAN, G., 1942. *Indian J. Phys.*, **16,** 77.
SWANN, W. F. G., 1946. *J. Franklin Inst.*, **242,** 167.
TEETER, C. E., Jr., 1946. *J. Acoust. Soc. Amer.*, **18,** 488.
TELFAIR, D., 1942. *J. Chem Phys.*, **10,** 167.
TELFAIR, D. and PIELEMEIER, W. H., 1942. *Rev. Sci. Instrum.*, **13,** 122.
TE-TCHAO, O., 1946 a. *C.R. Acad. Sci., Paris*, **222,** 1165.
TE-TCHAO, O., 1946 b. *C.R. Acad. Sci., Paris*, **222,** 1215.
THIEDE, H., 1941. *Akust. Z.*, **6,** 64.
THIEDE, H., 1943. *Akust. Z.*, **8,** 20.
TIMOSHENKO, S., 1934. *Theory of Elasticity* (New York: McGraw-Hill Book Co.), p. 9.
TISZA, L., 1938. *C.R. Acad. Sci., Paris*, **207,** 1035.
TISZA, L., 1940 a. *J. Phys. Radium* [8], **1,** 164.
TISZA, L., 1940 b. *J. Phys. Radium* [8], **1,** 350.
TISZA, L., 1942. *Phys. Rev.*, **61,** 531.

TISZA, L., 1947. *Phys. Rev.*, **71**, 477 A
TONKS, L., 1936. *Phys. Rev.*, **50**, 955.
TOWNSEND, C. D., 1946. *Products Finishing*, **10**, 72.
TRENT, H. M., 1945. *U.S. Dept. Commerce Publ. Bd. Report* **1672.**
TSIEN, H. and SCHAMBERG, R., 1946. *J. Acoust. Soc. Amer.*, **18**, 334.
VAN ITTERBEEK, A. and LAUWERS, L., 1946. *Physica*, **12**, 241.
VAN ITTERBEEK, A. and MARIËNS, P., 1937 a. *Physica*, **4**, 207.
VAN ITTERBEEK, A. and MARIËNS, P., 1937 b. *Physica*, **4**, 609.
VAN ITTERBEEK, A. and MARIËNS, P., 1940 a. *Physica*, **7**, 125.
VAN ITTERBEEK, A. and MARIËNS, P., 1940 b. *Physica*, **7**, 909.
VAN ITTERBEEK, A. and MARIËNS, P., 1940 c. *Physica*, **7**, 938.
VAN ITTERBEEK, A., and THYS, L. 1938 a. *Physica*, **5**. 640.
VAN ITTERBEEK, A. and THYS, L., 1938 b. *Physica*, **5**, 889.
VAN ITTERBEEK, A. and VANDONINCK, W., 1943. *Physica*, **10**, 481.
VAN ITTERBEEK, A. and VANDONINCK, W., 1946. *Proc. Phys. Soc.*. **58**, 515
VAN ITTERBEEK, A. and VERMAELEN, R., 1942. *Physcia*, **9**, 345.
VAN VLECK, J. H., 1932. *Theory of Electric and Magnetic Susceptibilities* (Oxford : Clarendon Press), pp. 269–271.
VENGEROV, M., 1946. *Nature, Lond.*, **158**, 28.
VENKATESWARAN, C. S., 1942. *Proc. Indian Acad. Sci., Bangalore*, **A15**, 362
WEBER, L. R. and GOEDER, F. P., 1942. *Phys. Rev.*, **61**, 94.
WEGEL, R. L. and WALTHER, H., 1935. *Physics*, **6**, 141.
WEIGLE, J., 1942. *Helv. phys. Acta*, **15**, 329.
WEIGLE, J. and BLEULER, K., 1942. *Helv. phys. Acta*, **15**, 445.
WEISSLER, A., 1947. *J. Chem. Phys.*, **15**, 210.
WEST, W. A., 1946. *Metals Technology*, **13**, T.P. 1993.
WILLARD, G. W., 1941. *J. Acoust. Soc. Amer.*, **12**, 438.
WILLARD, G. W., 1947 a. *J. Acoust. Soc. Amer.*, **19**, 235.
WILLARD, G. W., 1947 b. *Bell Laboratories Record*, **25**, 194.
WILLIAMS, E. J., 1933. *Proc. Roy. Soc.*, A, **139**, 163.
WILLIS, F. H., 1947. *J. Acoust. Soc. Amer.*, **19**, 242.
ZACHOVAL, L., 1939. *C.R. Acad. Sci., Paris*, **208**, 265.
ZENER, C., 1931. *Phys. Rev.*, **37**, 556.
ZENER, C., 1937. *Phys. Rev.*, **52**, 230.
ZENER, C., 1938. *Phys. Rev.*, **53**, 90.
ZENER, C., 1946. *Metals Technology*, **13**, T.P. 1992.
ZVETKOV, V., MINDLINA, A. and MAKAROV, G., 1946. *Acta Physicochim. U.R.S.S.*, **21**, 135.

LATENT IMAGE FORMATION IN PHOTOGRAPHIC SILVER HALIDE GELATINE EMULSIONS

By W. F. BERG,
Kodak Ltd., Harrow

§1. INTRODUCTION

The investigation of the photographic latent image represents a problem which has not so far been reported on in detail in these reviews, although certain aspects have been discussed in an article by Selwyn (1939). The reasons for this are not far to seek. The problem is not one of pure physics, since it lies in the border-land of physics and chemistry, both organic and inorganic, and of physical chemistry: some knowledge on all these matters or, alternatively, close team work is necessary for many of the problems involved to be tackled successfully. Furthermore, it is distinctly a problem of applied science, and the impetus for most of the research done derives from the commercial background. In consequence, work has tended to shift from the private and university laboratories to industrial establishments; requiring as it does a large amount of background experience it is probably fair to say that it is difficult if not impossible for a lone worker to make much progress in this field in the time available, say, for a Ph.D. thesis. In spite of all this, our problem cannot be said to have suffered from a lack of theories; in fact it is little exaggeration to say that fresh latent image theories used to be created at the rate of several per year, often on the strength of a few isolated experiments.

A further difficulty in the way of investigations on this subject is the fact that so far there has been no method which would allow the latent image to be detected except by development, which in itself obliterates the latent image. This fact lends a peculiar atmosphere to any attempts made to account for the formation of the latent image, all the evidence being of an indirect nature and difficult to prove by the simple and direct method of approach favoured by the physicist.

In spite of the unsatisfactory nature of the problem, by far the most important advance in understanding the mechanism of latent image formation is represented by a theory due to two theoretical physicists, Gurney and Mott (1938). Since its publication there have been but one or two new theories on the formation of the latent image. Their theory has been proved to be a valuable working hypothesis and caused a new outburst of research. We feel justified, therefore, in leaving out most of the earlier mechanisms suggested for the formation of the latent image, although this inevitably lends a certain bias to the present report.

Apart from any theories on the fundamental mechanism, it has become possible to an increasing extent to tie together various photographic effects as being due to common causes. Considerations and experiments, for example, on the distribution and dispersity of the latent image specks situated on a grain

can be made to account for many previously isolated experiences. Such realizations may be of greater value than a detailed picture of the fundamental mechanism since they will retain their validity even if the fundamental concepts require changing.

The author reported on the same problem in 1942 (Berg 1942). Since then research carried out during the war has been published or become available, although some work accumulated during the war still awaits publication. Space will allow rather fuller discussion of the problem than has previously been possible.

The main puzzle a theory on the mechanism of latent image formation has to solve is this:

Only quite a few quanta of light have to be absorbed by one of the small crystals or "grains" of silver halide in an emulsion to render this grain developable: the necessarily minute effect of these quanta will cause the whole grain to be reduced to metallic silver when the emulsion is treated with a solution containing certain reducing agents. The unexposed grains are not so affected. Our most important questions are:

(i) What is the change produced by the absorption of light.

(ii) What is the mechanism of development and, in particular, how does a developer distinguish between exposed and unexposed grains.

Under these main headings most or all photographic experience can be considered, although a large number of variables influence both the formation of the latent image and the action of the developer.

§2. A FEW PHOTOGRAPHIC FACTS AND TERMS

2.1. *General Information*

An ordinary photographic emulsion when coated on its support contains some 10^9 to 10^{12} crystals or grains of silver halide per square cm. held in suspension by gelatine. The crystals are between $\frac{1}{10}\,\mu$ and several μ in diameter, approximately $1\,\mu$ apart, and largely distributed at random; one reckons that there are some 40 layers of single grains on top of one another when considering the cross section of a coated and dried layer. The distribution in space of these grains cannot of course be completely random, both because their dimensions are comparable with their distances apart and because they carry on their surfaces surplus electrical charges tending to keep them separated.

These grains constitute the light-sensitive units which eventually build up the photographic image. In carrying out the photographic process the grains blacken if exposed, the dense parts of the image containing many, the light parts very few or none, of the grains. In the present Report our story is essentially that of the behaviour of the single grain, including, of course, its immediate surroundings, which will be seen to play a very important part. One of the difficulties in photographic research lies in the fact that as a rule it is impossible to study the individual grains; any conclusions drawn on their behaviour are derived from experiments involving very large numbers of grains.

The grains in an emulsion vary in size according to the way the emulsion is prepared. In the course of emulsion-making the silver halide is precipitated and then "ripened" in the presence of a silver halide solvent. During the

ripening process the grains increase in size, the larger grains growing at the expense of the smaller ones which have a higher solubility (Ostwald ripening). Much work has been done on measuring the grain size frequency distribution of emulsions and correlating it with the photographic properties. According to the method of preparation the emulsion grains are either hexagonal or triangular flat discs, the crystal faces being octagonal (111) faces in a cubic system, or cubes with often rounded corners and edges. The flat discs tend to be orientated preferentially with their large dimensions parallel to the support. This is because the gelatine on drying down on to the support shrinks some 10–15 times in a direction at right angles to the surface of the support whilst keeping its dimensions parallel to the support.

An exposure which is so small as to cause no alteration in a grain as far as can be detected by any physical methods will cause the grain to be reduced subsequently in the developer, a solution containing suitably chosen (usually organic) reducing agents. As a rule, after normal development grains are found to be

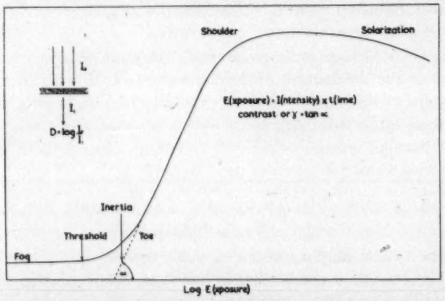


Figure 1. Diagrammatic representation of the characteristic curve of a photographic material to explain technical terms. (*After Berg*, 1942.)

either completely developed or not changed at all. In photographic theory the number of grains made developable in an emulsion is often considered, and has in fact been investigated directly by grain counting. This will allow the grain classes characterized by the different dimensions of the grains to be investigated separately. For most of the work, however, an optical measure is used which is chosen to be as nearly proportional as possible to the number of developed grains in an emulsion layer. It is clear, however, that the proportionality constant must vary from emulsion to emulsion and even with the development treatment. The optical measure which was found (see e.g. Trivelli 1946 a) to be within close limits proportional to the number of developed grains is the "density" which is defined as the logarithm to the base 10 of the ratio of the incident to the transmitted measuring light as illustrated in figure 1. The precise value of the density will depend to some extent on the optical set-up used, since the developed layer scatters as well as absorbs the light; the most satisfactory method is to use the "contact-printing density" by which literally all the light leaving the emulsion is measured.

The "characteristic curve" of a photographic material is a plot of the density against the logarithm of the exposure, i.e. the product of intensity and time—in other words, the total amount of light incident on a unit area of the emulsion surface. For theoretical purposes it would be much better to be able to use the amount of light actually absorbed by the emulsion grains, but this is often difficult if not impossible to determine for emulsions of normal coating thickness in which the light is scattered very considerably (see §4.6.3). The difficulty is increased by the fact that light is absorbed not only by the grains but by the gelatine as well. In spite of the limitations imposed thereby, most of the photographic information available is based on data obtained from experiments on normal materials. Both the shape and position of the characteristic curve of any one material depend on a large number of factors, some of which are extraneous to the material under investigation, and all of which must be specified if full information is to be conveyed. The curve varies with the nature of the developer used and its temperature and agitation, with the colour of the light and the nature of the modulator used to obtain the exposure steps, i.e. whether it was the intensity or the time that was altered, to mention only a few of the more important factors.

Various photographic terms are illustrated in figure 1. The characteristic curve is said to begin with a "toe" at which the density rises increasingly rapidly from the "fog" density obtained without exposure; this is followed by a more or less marked straight-line portion of the curve which turns over into the "shoulder". On further exposure the maximum density is reached; with some materials the density falls again, a phenomenon known as "solarization"; a further rise in the density is sometimes described as the second reversal of solarization. The slope ($\tan\alpha$) of the straight-line portion of the curve is called the γ (gamma). The "speed" or sensitivity of a photographic material determines the position of a certain point on the curve with respect to the log (exposure) axis. Various criteria are used as "speed points" such as the exposure necessary to produce a certain density on development, or the point at which the extension of the straight-line part of the curve cuts the fog level ("inertia" or "H and D" speed, after Hurter and Driffield). Other criteria adopted to characterize the speed of a material under practical working conditions have been proposed from time to time (see e.g. Jones and Nelson 1940) but need not be considered here.

2.2. *Photographic Effects*

There is an almost bewildering variety of photographic effects, most of which have the name of their discoverer attached to them. These effects are simply listed and briefly described here for convenience; of the more important ones a separate, more detailed account will be given later on:

2.2.1. *Dye-sensitization.* The spectral sensitivity of photographic materials is naturally limited to the region over which the silver halide absorbs. This region can be extended towards longer wavelengths by adding certain dyes to an emulsion; in this way orthochromatic, panchromatic and even infra-red sensitive materials are obtained.

2.2.2. *Desensitization by dyes* (Kornfeld 1938). Certain dyes when added to an emulsion will strongly repress its sensitivity to light. If applied in solution

after exposure, the existing latent image is not affected. This effect is often utilized in order to allow development of photographic materials under bright lighting conditions. A repression of the natural sensitivity is apparently always brought about when sensitizing dyes are used (Breido and Gorokhovski 1943).

2.2.3. *Reciprocity failure.* Bunsen's reciprocity law for photochemical reactions states that the amount of reaction product depends on the total amount of light absorbed and not on the individual values of intensity and time of exposure used to apply this amount of light. In this form the law applies to all single-stage photochemical reactions, but often breaks down when several stages are involved. This is the case with photographic materials, an effect described as failure of the reciprocity law or, shortly, as reciprocity failure. Figure 2 gives some examples of this effect and its variation with temperature during exposure. The usual representation of the effect is here followed by plotting the log (exposure) necessary to produce a certain density on development under standard conditions against the log (intensity). In this form the curve usually has a minimum or optimum where the exposure required is least. This minimum as a rule occurs at an intensity or time near that commonly used in photographic work, viz. between $\frac{1}{10}$ and $\frac{1}{100}$ sec. The drop in sensitivity at lower intensities and longer times is described as low-intensity reciprocity failure, or briefly low-intensity failure, that for higher intensities and shorter times as high-intensity failure.

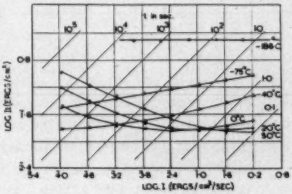


Figure 2. Reciprocity law failure curves of a photographic material at different temperatures. The log *It* necessary to produce a constant density is plotted against log *I*. The sloping lines are lines of constant time in seconds as marked along the top. (*After Mees*, 1942.)

2.2.4. *Schwarzschild or intermittency effect.* As Schwarzschild effect has been described the fact that an intermittent exposure often produces a different result from an exposure of the same duration and same average intensity. This effect is linked with the occurrence of reciprocity law failure as will become clear below.

2.2.5. *Kron effect* (Kron 1913, Strock 1938). The fact that the position of the optimum of the reciprocity failure curve varies with the degree of development used, shifting towards higher intensities and shorter times as the duration of development is increased, is sometimes described as the Kron effect.

2.2.6. *Latent image intensification.* Latent image intensification, or "latensification", is a method by which the latent image in an emulsion can be built up or intensified so that in effect a larger proportion of the exposed grains is made developable. The effect was originally obtained by treating an exposed emulsion with hydrogen peroxide (Wightman and Quirk 1927), and has more recently been studied when brought about by a post-exposure to light of very low intensity or by treatment in mercury vapour (Moore 1941, Sheppard, Vanselow and Quirk 1945, Burton and Berg 1946, Jausseran 1943, Dersch and Dürr 1937, Rua 1942).

2.2.7. *Hypersensitization.* The photographic speed of a given product can be increased by various treatments applied before exposure and usually described as hypersensitization. The effect can be brought about by treating the material with mercury vapour (Dersch and Dürr 1937) or by applying a pre-exposure to light of short duration (Burton and Berg 1946, Weinland 1927, 1928, Webb and Evans 1940b, Jausseran 1943, Rua 1942). Basically, latent image intensification and hypersensitization by exposures to light are the same effect. In each case an exposure of short duration is followed by one of long duration; when intensifying, we are interested in recording the former, when hypersensitizing, the latter. Infra-red sensitive emulsions can be hypersensitized by treatment with ammonia and other chemicals (Clark 1939). In each case the material becomes either unstable so as to deteriorate on keeping, or at least somewhat foggy in the course of the treatment.

2.2.8. *Herschel effect.* With certain photographic materials which are not themselves sensitive to red or infra-red light a post-exposure (before development) to light of this quality will cause a decrease in the density, produced by a white light exposure, due, apparently, to a destruction of the latent image. Originally, however, the term "Herschel effect" referred to a similar reduction in density brought about by a red-light post-exposure on a print-out image.

2.2.9. *Weigert effects.* Two different, although related, effects have been described by Weigert (Freundlich 1936, Cameron and Taylor 1934). One refers to the "print-out" image (see §3.1) obtained by direct photolysis of silver halide emulsions, the other to exposures merely producing a latent image. If an exposure is given to a polarized white light and followed by an exposure to coloured and polarized light, a dichroic print-out image is obtained varying with the colour of the second exposure. If the first exposure is weaker, so as to produce a latent image but no print-out image, and is exposed to polarized light, the developed image turns out to be dichroic. These phenomena only occur in very fine-grain emulsions, in which the light is not depolarized by scattering. Too little attention appears to have been paid to these interesting effects.

2.2.10. *Seebeck effect* (see Zenker 1868). The Seebeck effect or "photo-adaptation" is the fact that in materials containing mainly silver chloride the colour of the print-out image tends to be altered to that of the light used in a second exposure. These print-out effects on chloride emulsions have not been studied for many years.

2.2.11. *Clayden effect* (see Wood 1903). The Clayden effect is the desensitization which follows an exposure to light of very high intensity and short duration when the testing exposure is one of lower intensity and longer duration. Thus a night photograph of a stroke of lightning may appear reversed on the uniform background produced by the general illumination of the scene acting for a longer time.

2.2.12. *Villard effect* (see Wood 1903). A similar desensitization is brought about when the first exposure is to x rays, the subsequent testing exposure to ordinary light.

2.2.13. *Pressure effect* (see Wood 1903). Ordinary hydraulic pressure appears to have little or no photographic effect. When the force applied is not homogeneous, however, a latent image may be produced. This is accompanied

by a desensitization towards an exposure to light following the mechanical distortion. It is said that in the order, pressure, X rays, high-intensity light, low-intensity light, any of the previous agents will produce a desensitization towards any of the subsequent ones.

2.2.14. *Sabattier reversal* (Sabattier 1927). A photographic material, if pre-exposed and developed for a short time, gives a positive image with a second exposure; this leads to peculiar border effects in the not infrequent event of the light being switched on accidentally during development. Optical screening by the image partly developed after the first exposure is to a large measure responsible for the reversal (Marriage 1937), but it has also been shown that the grains are desensitized during the first development (Stevens and Norrish 1938), since chemical "fogging agents" like sodium arsenite can produce the effect. The mechanism of this desensitization is not clear. It has been suggested that silver is transferred from developing to neighbouring grains, which latter are thereby supposed to become desensitized. The transfer has been confirmed (Stevens and Norrish 1938), but no clear mechanism has been suggested as to why this should cause a desensitization.

2.2.15. *Albert reversal* (Albert 1899, Stevens 1939). A similar reversal is obtained when a material is exposed and then treated with chromic acid before the second exposure and development. This effect is not of practical importance but has considerable theoretical significance.

2.2.16. *Dye reversal* (see e.g. Carroll and Kretchman 1933). This reversal is obtained in a way similar to the two previous ones except that the chemical treatment following the first exposure consists in bathing the material in a solution of certain dyes. The second exposure will lead to a reversed or positive image if the colour of the light is such as to be absorbed partly or completely by the dye. The mechanism of this reversal is thought to be similar to that of the Herschel effect; it is therefore often described as a sensitized Herschel effect, the sensitization referring to the fact that different wavelengths of light are capable of producing the reversal. The mechanism of the effect is not clear.

2.2.17. *Waterhouse reversal* (Waterhouse 1890, Rawling 1926). A reversed (positive) image can be obtained when thiocarbamides or other labile sulphur compounds are added to a developer. The mechanism of this reversal will be discussed below (§ 5.3).

2.2.18. *Sterry effect* (Sterry 1904). The contrast of a material can be lowered by treating the exposed material before development with certain chromates.

2.2.19. *Russell effect* (see Bullock 1937). Certain metals, mainly zinc, and other chemicals and substances of organic origin, are capable of both forming a latent image and of desensitizing the material towards a subsequent exposure to light. It is said that this effect is often due to the spontaneous production of hydrogen peroxide.

2.2.20. *Ross or gelatine effect* (Ross 1924). In the course of development a small image may be found to have contracted due to the tanning action of the development products on the gelatine. This effect is of importance in astronomical work; when there are two star images close together they may appear to be attracted to one another.

2.2.21. *Eberhard, Kostinski, Nachbar effects* (Ross 1924, Wolfe and Burrows 1947). The value of the density in an exposed and processed photographic material is often not independent of the density values of the surrounding areas. Thus if a heavy density is surrounded by an area of lower density, the value of the lower density may be decreased near the border between the two, and its own density may be raised. This may lead to an apparent repulsion of two neighbouring star images (Kostinski) or to a bright line (Mackie line) surrounding a dense area in a negative. These effects are believed to be due merely to local changes in concentration of the developer and development products and are diminished by proper agitation of the developer solution.

2.2.22. *Purkinje effect.* In accordance with the similarly named visual effect, the photographic Purkinje effect consists in a change in the gamma (see figure 1) of the photographic material with changing wavelength of the exposing light.

2.2.23. *Lainer effect* (Lainer 1891, Lüppo-Cramer 1927). Treatment of a photographic material after exposure with an appropriately chosen solution containing iodide ions may lead to an apparent increase in sensitivity of a photographic material, brought about by an increase in the rate of development.

2.2.24. *Capri blue effect* (Lüppo-Cramer 1940). Sensitivity and contrast of a coarse grain emulsion which has not been chemically sensitized can be increased by certain dyes, which do not necessarily extend the spectral region over which the material is sensitive.

§3. NATURE OF THE LATENT IMAGE

3.1. *Chemical*

On exposure, working on the basis of quantum equivalence, approximately 1 in 10^9 ion pairs in a silver halide grain can at the most be affected by light when a latent image is formed near the threshold of the characteristic curve. It is this very small proportion of changed ions which makes it so difficult to decide what is the true nature of the latent image. When exposures are increased by a factor of 10^7 to 10^8 the photographic material "prints out"; it darkens visibly due to a process of reduction of the emulsion grains partly or wholly to metallic silver and halogen. With an ordinary silver bromide emulsion this process cannot proceed further than to cause a just visible change, when a saturation density is reached; that is because the bromine released will attack any more silver formed. If, however, a halogen acceptor is added to the emulsion, printing out will proceed until the grains are completely reduced and the density becomes similar to that obtained on developing a material which has received an exposure just sufficient to form a latent image on most or all of the grains. By very sensitive absorption measurements this process of printing out has been followed towards smaller exposures back into the solarization region of photographic materials, the nature of the products of exposure not varying throughout the range tested (van Krefeld and Jurriens 1937), but it has not been possible to follow the process into the latent image region itself by purely physical methods.

From these facts, however, photographic workers have early deduced that the latent image itself consists of metallic silver in intimate contact with the parent crystal (credited to Guthrie; see Abney 1882). This concept is now almost

generally accepted, but for a long time other theories were widely discussed, such as the suggestion that the latent image is a sub-halide (Abney 1882, Carey Lea 1918). This suggestion was based on the finding that the discoloration formed in certain silver chloride preparations by absorption of light could not be bleached by oxidizing agents. It was deduced that these "photo-halides" represented a new chemical substance. At present it is thought that the photo-halides are an adsorption complex between silver and silver halide, some of the silver being in the interior of the crystals and thus protected from the bleaching agent acting from the surface of the grains. Eventually it was demonstrated conclusively that the photo-halides consist of silver and silver halide, and that, on photolysis, silver and halogen are the only products (Baur 1903, Reinders 1911). Evidence bearing in a similar direction is found in the fact that the kinetic reaction curve of the bromination of silver has no kink (Hartung 1924), which means that a single chemical substance is formed, thus excluding the possibility of formation of sub-halides. A large amount of work was done to identify the latent image substance by its chemical behaviour, which is summarized in the book by Bullock (1927). Finally, crystal structure investigations by x-ray (Koch and Vogler 1925) and electron diffraction (Trillat and Merigoux 1936) methods of exposed silver halide layers show that the formation of silver on the grains was the only result of exposure (apart from the halogen liberated). In all these investigations exposure had to be carried beyond the level necessary to obtain a latent image in the material, except in the case of some exceptionally fine-grain emulsions. To assume, therefore, that the latent image itself is metallic silver constitutes an extrapolation, although no argument has been advanced why this extrapolation should not be made.

Another suggestion sometimes made is that the latent image constitutes a break in the electrical double layer which normally surrounds the emulsion grains (Schwarz and Urbach 1932 a, b, Schwarz 1933, Evans and Hanson 1937). Various results of photographic experience contradict this point of view. Thus, it is possible to produce a latent image in a grain and then to change the nature of the parent crystal by a double decomposition process, say from silver bromide to silver iodide (Renwick 1920). This process always occurs when a chloride (paper) emulsion is developed in a solution containing bromide ions. The latent image withstands this treatment; it would be difficult to imagine how anything so ephemeral as a local variation in surface charge of a grain could survive this drastic procedure. Similar treatments can be applied to reveal latent images formed in other substances such as thallous bromide, in which exposure produces a (thallium) latent image. On treatment with silver ions this is changed into a silver latent image, the thallous bromide grain into a silver halide crystal or crystalline aggregate (Farrer 1936, Ritchie and Thom 1946). Furthermore, a latent image can be revealed by a process known as "physical development", in which the emulsion after exposure is first fixed out and then treated with an appropriate plating bath which deposits silver, mercury, or copper on any nuclei present. Again the latent image so revealed must have a reasonably substantial material existence. These findings seem to eliminate the possibility of the latent image being a mere physical change in the grain surface.

Similar considerations would seem to apply to the ingenious suggestion

recently put forward by Huggins (1943). He has demonstrated that silver bromide crystals are likely to occur in two different crystalline forms, the common rock salt (B1), and the zinc sulphide type (B3) in which silver iodide exists at room temperature. The lattice energies for these two structures of silver bromide are very similar indeed, so far as they can be calculated. In the B3 structure there are wider gaps than in the B1; these gaps may accommodate either surplus positive or negative ions. The latent image is thought to consist in a region in which the B1 structure has been changed into B3 by the action of photo-electrons. No experimental evidence has been presented to prove the correctness of this model; it would appear possible to obtain this proof by x-ray diffraction methods applied to developing silver halide crystals, since in these the B3 regions are supposed to extend rapidly at the expense of the B1 regions. For very long exposures Huggins considers that the B3 regions break down into metallic silver as obtained in printing-out.

3.2. *Physical*

Interesting evidence which is likely to have a direct bearing on the latent image problem has been obtained by the work of Hilsch and Pohl and their school (Hilsch and Pohl 1930, Löhle 1933, see also de Boer 1935) on large crystals of silver halides. These workers argue that the main reason why the latent image in an emulsion cannot be investigated by physical methods is the very small thickness of silver halide that can be utilized owing to the scattering of light in emulsion layers. A much larger thickness of silver halide can be tested when single crystals are used. In single crystals of silver chloride and silver bromide a discoloration is obtained by exposure to light of the order of that producing a latent image in a normal photographic material. Absorption of blue light by silver bromide crystals produces a new absorption band with a maximum at 690 mμ. Details of this process will be discussed below. Here we note that absorption of light of wavelengths falling into the new absorption band causes destruction of the band around the wavelength used, leaving parts of the band away from this wavelength unaffected. The band is thus due not to homogeneous material but to particles of a colloidal nature. The position of the band depends on the size of the particles, and absorption of light causes the particles to be destroyed or dissipated. An exactly parallel effect obtains in a photographic emulsion in which red light causes a destruction of the photographic latent image (Herschel effect, see § 2.2.8). The wavelength sensitivity of the Herschel effect corresponds closely to the absorption band in discoloured silver halide crystals (Carroll and Kretchman 1933, Bartelt and Klug 1934).

Interesting energy considerations by Bodenstein (1941) indicate that photolysis of silver halides cannot proceed unless the silver atoms formed condense into metallic specks. The reason is that the energy of the light quanta capable of producing photolysis is insufficient, on balance, unless the condensation energy of silver is brought into the picture. Details of Bodenstein's considerations have been criticized by Stasiw and Teltow (1944), but the basic argument appears to be unaffected.

Other evidence of photographic origin also indicates that the latent image consists of small specks of silver distributed over the surface of the grain.

Further evidence suggests that these specks can occur in the interior of the grains as well (§ 5.4). Development has been shown to begin on one or a few isolated points on the grain (Hodgson 1917); these points, the so-called development centres, have been identified with the seat of the latent image specks. In his classical experiments, Svedberg (1922) optically projected the partially developed grains so that the image of each grain fitted a circle of constant size, and marked in the circle the position of the development centres. If the centres were distributed throughout the grains, there should be a larger number of centres in the middle of the circle than near the circumference; if the centres were seated on the grain surface, they should appear to be concentrated near the edge. The results left no doubt that the centres were distributed haphazardly over the grain surfaces. We now know that this merely indicates that a normal developer can utilize only those latent image specks with which it is in immediate contact.

The mechanism by which silver produced by absorption of light all over the grain aggregates on a few points in and on the crystals is the most important part of any theory on latent image formation.

One would conclude that all the experimental evidence available agrees with the conception that the latent image as formed by exposing silver halides to light, certain other radiations and chemical agents, consists of small specks of colloidal silver; during its formation halogen, originally in atomic form, must be liberated and has been determined by chemical methods (§ 4.6).

§ 4. PROPERTIES OF SILVER HALIDES

In order to understand the mechanism of photolysis of silver halides it is essential to review as closely as possible the knowledge now available about silver halide crystals. In doing this it will be found that our knowledge of the material in bulk is much more satisfactory than that of the very small crystals involved in photographic emulsions. With these latter, the surfaces play a very important part, and for obvious reasons are difficult to investigate.

4.1. *Composition of Emulsion Grains*

In normal photographic negative emulsions the bulk of the silver halide is silver bromide. Usually a small amount (up to about 7 %) of silver iodide is added so as to form mixed crystals of bromide and iodide (Wilsey 1921, 1925), still crystallizing in the cubic system, but with a slightly larger lattice constant than that of pure silver bromide; there is a slight irregular distortion of the structure (Brentano and Baxter 1940). If larger amounts of iodide are added a solid solution of iodide and bromide is no more possible, the iodide itself crystallizing in the B3 (cubic) or B4 (hexagonal) system. The effect of the iodide on the behaviour of photographic emulsions is not properly understood; it is certain that in the presence of iodide it is more difficult to grow the emulsion grains to a large size. Silver bromide or iodo-bromide emulsions have been studied more extensively than silver chloride emulsions, which may contain pure silver chloride crystals or mixtures of chloride and bromide in varying proportions. Since both chloride and bromide crystallize in the same system, mixed crystals of any proportions can be obtained (Wilsey 1921, 1925).

4.2. *Binding and other Energies*

Various energies applying to silver halide crystals have been collected in table 1, most of them from the work by Mayer (1933). It will be seen later (§ 4.4)

Table 1. Energy of silver halide crystals (kcal.)

	AgCl	AgBr	AgI
Lattice energy calculated electrostatically (Mayer 1933)	203	197	190
By Born cycle (Mayer 1933)	206	202	199
Hydration energy (Mees 1942 a) ..	192	183	170
Heat of solution (Mees 1942 a) ..	−16	−20	−27
Energy of formation of interstitial Ag^+ (Koch and Wagner 1937)	25	20·2	—
Activation energy for mobility of interstitial Ag^+ (Koch and Wagner 1937) ..	6·5	8·2	—

that the work involved in producing silver ions in interstitial positions is sufficiently low to ensure even at room temperature an appreciable ionic conductivity (Koch and Wagner 1937, Jost 1938).

4.3. *Surface Layers*

In many instances the behaviour of a photographic emulsion is largely controlled by the state of the surface layer of the grains. Since emulsions have very large specific areas (Leermakers, Carroll and Staud 1937 b), between 400 and 1000 cm^2/cm^3, considerable amounts of chemicals can be adsorbed to the grains, the most striking example being perhaps that of the sensitizing dyes to be discussed in more detail later (§ 4.6.6.). All ordinary emulsions are prepared with a surplus of halide ions; in consequence the grain surfaces carry a surplus of halide ions and are thus negatively charged. The evidence on the amount of halide ion so adsorbed, and therefore on the strength of the electrical double layer surrounding the grains, is conflicting. It has been stated that up to 10 layers of bromide ions may surround a grain, a suggestion which is obviously impossible (Eggert and Noddak 1927 b). The difficulty in investigating the halide layer is due to the strong adsorption of the gelatine in which the grains are suspended, since the gelatine is amphoteric. The effective charge on the grains is therefore determined not by the bromide ions but by the hydrogen ion concentration in the emulsions (see also § 6.2.2).

Most of the evidence on adsorbed layers has been obtained on silver halide sols in the absence of gelatine. Surface layers of hydroxyl ions, silver ions and other chemicals have been investigated, and cause changes in the spectral sensitivity of the silver halide crystals, as will be discussed below (§ 4.6.4). In considering the formation of chemi-sorbed layers it is valuable to remember Paneth's rule, viz. that if there are two competing substances surrounding the grain, that one will be adsorbed preferentially which would tend to form a less soluble compound with one or the other of the ions in the crystal. Apart from that, mass-action considerations apply, i.e. the ion or molecule present in larger concentrations will tend to get adsorbed.

4.4. *Ionic Processes in Silver Halide Crystals*

The high ionic conductivity in silver halide crystals is one of the properties which distinguishes them from most other ionic crystals. This conductivity is of importance to all considerations on the photolysis of silver halides; as will be discussed below, it can occur both by ions in interstitial positions and by the holes left in the structure when interstitial ions are formed.

4.4.1. *Interstitial ions* (Koch and Wagner 1937, Jost 1938). The interstitial silver ions always formed in silver halide crystals in thermal equilibrium must be imagined to occupy positions in the centre of the unit cube of the structure. Some distortion of the structure occurs and thus a formation energy is involved. The bromide ions are too large to be able to occupy interstitial positions. The interstitial silver ions are mobile, although requiring a certain activation energy (see table 1, p. 000). The concentration of interstitial ions is given by a formula of the type $C = C_0 \exp(\frac{1}{2}W/kT)$ and is of the order of 10^{-6} mol. for silver bromide and 10^{-8} mol for silver chloride at room temperature.

4.4.2. *Holes* (Koch and Wagner 1937, Jost 1938). The hole left in the structure when an interstitial ion is formed has been shown to be itself mobile, the mechanism being a replacement; a neighbouring ion jumps into the hole, leaving a hole at the place whence it came (mechanism 3 in figure 3). The activation energy of this process is of the same order of magnitude as that for the movement of the interstitial ions. Holes and interstitial ions, therefore, contribute similar amounts to the overall electrolytic conductivity.

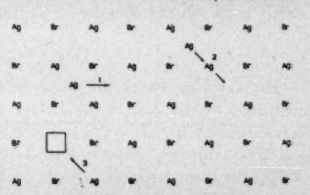


Figure 3. Mechanism of ionic conductivity in silver halides (bromide). The Ag and Br ions are arranged as in a cubic face of the crystal. In (1) an Ag ion moves from one interstitial position to another, which must be imagined to be situated one half of a lattice distance below or above the paper plane; in (2) a silver ion in a normal position moves into an interstitial position and is replaced by one coming from an interstitial position; (3) is the replacement mechanism by which the "holes" move. (*After Berg*, 1943 b.)

The concentrations of both interstitial ions and holes in a crystal may be affected by the presence of impurities, such as bi-valent ions, which according to their charge are balanced by surplus silver ions or holes. Furthermore, at low temperatures the crystal may not be in thermal equilibrium and may thus contain a higher concentration of interstitial ions or holes than might be expected; this probably applies at room temperature, where the temperature coefficient of the conductivity seems to depend on the mobility of the ions only (Toy and Harrison 1930, Lehfeldt 1935, Berg 1940).

Various modes of movement of interstitial ions have been illustrated in figure 3. The figure represents a cubic face of a silver bromide crystal; the interstitial ions must be imagined to be one half of the silver-bromide distance above or below the plane of the paper.

It has been estimated that at room temperature it takes 10^{-3} to 10^{-6} seconds to form a thermal equilibrium concentration of interstitial ions in silver bromide

and between 10^{-2} and 10^{-1} seconds for silver chloride, assuming that there are no interstitial ions at the beginning (Berg 1943).

4.4.3. *Small crystals.* The values for the concentration of ions and holes in the small crystals which exist in a photographic emulsion are likely to differ considerably from those in the bulk material. Grimley and Mott (1947) have investigated theoretically the conditions existing at the interface between silver bromide and an electrolyte or metallic silver respectively. The calculations show that adsorption in the normal meaning of the term of bromide or silver ions to a silver halide surface is unlikely to occur. The crystal is negatively charged because near its surface there exists a surplus of vacant silver lattice points, and in the electrolyte surrounding it there is a surplus of silver ions. As a result, the ionic conductivity of the crystal near the surface may be several orders of magnitude higher than that in the bulk. The whole of an emulsion grain must be considered as being "near the surface" in this sense.

Similarly, near the silver bromide-silver interface, there is a stoichiometric surplus of vacant silver lattice points. Thus a thin layer of silver halide on metallic silver will have a considerably higher ionic conductivitiy than a large crystal.

Evidence bearing on this picture has been presented by Zimens (1946), who has studied the rate of interchange of silver ions between an emulsion and a solution containing radioactive silver ions. The rate of exchange is so high that it sets a lower limit for the ionic conductivity inside the crystals which must be at least ten times as high as that for bulk bromide. Unfortunately, the validity of this experiment is doubtful, since the combination of silver ions with gelatine (Carroll and Hubbard 1931 b) was not taken into consideration; we do not know whether the radioactive silver ions are taken up by the gelatine or the emulsion crystals.

Similar methods were used by Kolthoff and O'Brien (1939) to study the rate of interchange of radioactive bromide ions between precipitates of silver bromide and an electrolyte. They obtain measurable interchange under conditions where Zimens (1946) claims that there is no interchange at all.

Shapiro and Kolthoff (1947) have studied the conductivity of silver bromide as a function of temperature and pressure and have concluded that surfaces play an important part in the mechanism of conduction by producing vacant lattice points. Lack of space prevents further discussion of the interesting work by Kolthoff and his school.

4.5. *Electronic Processes*

4.5.1. *Photoconductivity.* A silver halide crystal kept in the dark at sufficiently low temperature is an electric insulator. Absorption of light by the crystal produces a conductivity of electronic nature (Wilson 1907, Toy 1928, Toy and Harrison 1930, Tubandt 1932, Kameyana and Fukamoto 1939, Kameyana and Mizuta 1939). Each quantum of light absorbed is capable of producing an electron which can be drawn across the crystal to the anode if an electric field of sufficient strength is applied (Lehfeldt 1935). For insufficient field strength the current for constant illumination is decreased indicating that some of the electrons are trapped on the way. The nature of these traps is unknown;

they are likely to consist in physical imperfections of the crystals since the amount of trapping depends to a large extent on the thermal history of the material; a well-annealed crystal has fewer traps than a chilled one.

Experiments on this so-called photoconduction are as a rule carried out at low temperatures in order to avoid the disturbing ionic conductivity occurring at room temperature. Photoconduction, however, does not depend on temperature over a very wide range. It is said that conduction decreases only for temperatures below 20 °K. (Lehfeldt 1935, Pohl 1937).

The electrons carrying the photo-current are considered to be removed from the halogen ions in the crystal lattice and raised (in energy) into the so-called conduction levels (figure 4). In the conduction levels the electron moves about with thermal energy, its path velocity being of the order of 10^7 cm/sec.; its diffusion coefficient is of the order of 1 cm^2/sec. A crystal of silver halide containing colloidal metal particles or particles of silver sulphide exhibits the effect of trapping of electrons to a much greater degree than a pure crystal (Pohl 1937). This means that the conduction levels of silver or silver sulphide are below those of the silver halides (Gurney and Mott 1938), so that electrons passing from the silver halide into a silver speck cannot easily return. It appears that a given crystal cannot contain more than a certain number of these impurity trapping centres. It has been shown that on sufficient exposure to light the effect of trapping diminishes again, indicating that all the available traps are now filled and the mean free path of electrons is increased (Pohl 1937, Stasiw and Teltow 1941 a).

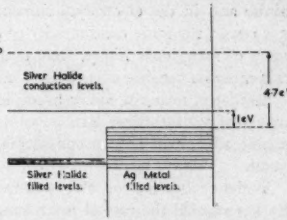


Figure 4. Electron levels in silver halide and colloidal silver. Absorption of light by the crystal raises an electron from the filled to the conduction levels; light of longer wavelengths may eject an electron from the silver to the conduction levels or from the filled silver halide levels to the silver.

Recent Russian work (Barshchevksy 1946, see also Toy 1928) deals with the photoelectric effect as a function of wavelength for very thin layers of silver halides. The effect depends on thickness and wavelength, since for strongly absorbed light absorption is confined to a very thin layer and recombination of electrons and halogen atoms plays a large part. Light of longer wavelength is absorbed at the edge of the absorption band and thus penetrates the crystal more deeply, causing a stronger photoelectric effect. The effect decreases with increasing time of illumination on very thin layers, which decompose under the action of light. With thick layers, the effect increases with increasing time of illumination, the maximum at the same time shifting towards longer wavelengths. This is explained by the production of colloidal silver particles; there seems to be no evidence in this paper for the trapping of electrons on these particles. Recently, photoconductivity has been demonstrated to occur in actual emulsion layers at room temperature and under normal exposure conditions (West and Carroll 1947 a). The photo-currents are proportional to the intensity of light

for the low intensities necessary to form a latent image, and tend towards a square-root dependence at high intensities. This has been interpreted to mean that the electrons are disposed of largely by trapping at low intensities, and largely by recombination with bromine atoms at high intensities of light. Dye-sensitization of photoconductivity was also found and will be discussed later (§ 4.6.6).

4.5.2. *Electron levels in silver bromide.* From data on the photoconductivity in silver halides and the photoelectric effect of silver, it is possible to draw up a picture of the electron levels in a silver bromide crystal (figure 4). The basis for this is the long-wavelength limit of the photoelectric effect on metallic silver which occurs at a wavelength of 330 mμ and is therefore 4·7 ev. (see e.g. Hammer 1922). When colloidal silver particles occur in silver bromide crystals, absorption of light by the particles of a wavelength around 700 mμ leads to an emission of an electron into the conduction band of the crystal. Although the long-wavelength limit of this emission is not accurately known, the difference in the energy levels for an electron between silver and the conduction band of the silver bromide crystal must be of the order of 1 ev. One aspect has been left out of this consideration, viz. that a small speck of silver is likely to represent a state of higher energy for an electron than a large piece of metallic silver. Whilst this effect undoubtedly exists it is not thought to affect materially the picture in figure 4.

4.5.3. *Electric capacity of colloidal particles in silver halides.* For many considerations it is interesting to consider how many electrons can be held trapped at any one time in one of the colloidal particles which can act as trapping centres. It has been shown (Berg 1943) that no more than a single electron can occupy one of these traps, the potential rise for a single electron being between 4 and 0·25 volt for specks containing between 4 and 10 000 atoms or molecules. Thus a single electron charges a colloidal particle up to a voltage to surmount which an energy is required considerably in excess of the average thermal energy available ($3/2kT = 0{\cdot}04$ ev.). Thus it is very unlikely for more than a single surplus electron to be found at any one instant on any colloidal particle or other kind of electron trap.

4.5.4. *Halogen atoms.* When an electron has been released from a halogen ion in the crystal, a halogen atom remains. Unless removed, this atom can serve as a trapping centre for electrons which are thus lost by recombination. When recombination does not occur, however, the halogen atom must be freely mobile in the crystal by a mechanism most easily imagined as an electron replacement process; an electron from a neighbouring halogen ion moves on to the atom, turning it into an ion and leaving an atom whence it came (Berg 1943). Because of the wave-mechanical tunnel effect no activation energy is involved in this process; in other words, the bromine atom should be as freely mobile as an electron in the conduction band but behaving like a positively charged electron. No experimental evidence has been produced to support this picture although there is little doubt about its correctness. It has been shown, however, that halogens can effectively penetrate into crystals and oxidize colloidal silver there (Stevens 1942 and unpublished experiments, Stasiw and Teltow 1944, Berg 1948).

It will be seen below (§ 5.1) that in order to produce a silver speck on a silver halide crystal during the process of latent image formation, it is essential for the

halogen atom to be held at least until the silver speck is formed. This is because the halogen atom and the electron will tend to draw together again, being effectively of opposite charge and highly mobile. There is evidence that in a photographic emulsion at least some of the halogen leaves the crystal and is taken up by the surrounding gelatine. One could imagine this process to occur with high efficiency if it were assumed that absorption of light occurred mostly near the surface of the silver halide grains. It is not necessary, however, to assume this since the movement of the halogen atom, even under the influence of an electrical field, is by a diffusion process; thus if a halogen atom were formed inside the crystal it is likely to have many opportunities of touching the surface of the crystal before reaching a free electron. On contact with the surrounding gelatine it would then be bound chemically and thus made harmless (Berg 1943). It has been thought for a long time that the chemical reaction between the released halogen and gelatine was sufficiently strong to ensure that the halogen was made harmless. Recent evidence, however, suggests (Berg 1948) that bromine released during exposure is capable of migrating distances of several millimetres laterally through the emulsion layer without being permanently bound. Effects produced by this bromine are pronounced at low intensities of light at which heavy exposures are needed to produce a latent image on an appreciable number of grains. Those grains, however, on which a latent image is formed in the early stages of such a heavy exposure are naturally subjected to a far-reaching photolysis and thus constitute strong sources of bromine. Certain emulsions exhibit this bromine effect more than others, but it appears that the halogen acceptance of gelatine has been overrated in the past.

A detailed discussion on the behaviour of atomic halogen in the photolysis of silver halides has recently been given by Bagdassarian (1944 b).

4.6. *Absorption of Light*

4.6.1. *Absorption by pure silver halides.* The absorption of light by silver bromide and chloride differs from that of most other ionic crystals in one very important respect. Instead of the sharply defined absorption bands found, for example with alkali halides, they exhibit an ill-defined absorption band falling slowly towards longer wavelengths (Fesefeldt and Gyulai 1929). This gradual fall is often described as the "tail" of the absorption band and has been accounted for by assuming that with silver halides the surface ions are considerably more loosely bound than those in the body of the crystal (Herzfeld 1923).

The assumption that the surface ions are loosely bound is further supported by the fact that the absorption of crystals, or at least the absorption tail, is easily changed by impurities. This will be discussed below, but it is clear that the long-wavelength absorption of silver halides cannot be regarded as typical for the basic material. A large single "real" crystal has been regarded as being built up like a house of bricks of very small "ideal" mosaic crystals (see Ott 1928). On all the interfaces between the individual ideal crystals the more loosely bound ions occur and thus form an appreciable fraction of the total number of ions, being independent of the size of the crystal. This independence, however, breaks down when the dimensions of the crystals are small, as those of emulsion grains, which may be regarded as being nearly perfect crystals in which only

genuine surface ions are loosely bound. With emulsion grains, then, the surface plays an increasingly important part as the grain size decreases.

The existence of this absorption tail is of crucial importance for the photographic process. If one imagines an emulsion made up from one of the alkali halides with a well-defined absorption band, one can see that all the light which the grains are capable of absorbing will be so absorbed in the first layer. With silver halides, light of a wide range of wavelengths is absorbed to some extent, but not too heavily, and is thus capable of penetrating into the depth of a photographic emulsion layer and of producing there a latent image.

The shapes of the absorption curves of silver chloride and silver bromide are similar in character, the chloride absorbing practically no visible light, the bromide essentially blue light, thus appearing pale yellow in colour. Silver iodide is a rather special case, the absorption band displaying a sharp subsidiary peak at its long wavelength end, beyond which it falls sharply. This subsidiary peak is matched by a sharp rise in sensitivity of the material (Huse and Meulendyke 1926).

4.6.2. *Absorption by mixed silver halides.* Partly because of the limited range of wavelengths over which the pure silver halides absorb light, and partly for emulsion-technical reasons, mixed crystals of silver halides are widely used in photographic emulsions. Mixed crystals of silver chloride and silver bromide absorb intermediately between the two pure substances (Breido and Gorokhovski 1944). The behaviour of mixed crystals of bromide and iodide is more intriguing. Up to approximately 30 mol. per cent, the addition of iodide to silver bromide causes an extension of the spectral sensitivity, i.e. necessarily one of absorption, well beyond that of either of the pure materials (Eggert and Biltz 1940). Above this percentage the system no more constitutes a single phase and the specific absorption band of the silver iodide begins to show up (Breido and Gorokhovski 1944). It may be pointed out here, that the increase in the range of wavelengths absorbed caused by adding iodide is coupled with an effective increase in sensitivity, which, however, considerably exceeds that due directly to the larger amount of light absorbed by the system.

4.6.3. *Absorption by emulsions.* Whereas it is a simple matter to investigate the spectral absorption of a single crystal, it becomes very complicated to assess the absorption of a photographic emulsion, because emulsion grains scatter light to a disturbing degree. The path length of light on its passage through an emulsion cannot be measured and the absorption curve is falsified in much the same way as that produced by an integrating sphere covered with a coloured material; where absorption is weak it becomes exaggerated and thus the absorption tail is even more pronounced.

Clear data on the genuine absorption of photographic emulsions can only be obtained by the use of "single layer" plates, viz. materials coated so thinly as to avoid the effect of scattered light altogether. By this method it has been shown that the absorption of light, the photoconductivity, and the spectral sensitivity of the emulsion coincide to a satisfactory degree, as they should if these processes are linked together (Toy 1928, Toy and Harrison 1930).

Photographic emulsions exhibit an absorption tail beyond that known from single crystal measurements; this appears to be specific to the silver halide-gelatine system. With silver bromide emulsions the absorption tail exhibits

a pronounced kink at 460 mμ, as judged by their spectral sensitivity; for longer wavelengths the sensitivity tail falls less slowly than for shorter wavelengths (Eggert and Kleinschrod 1940). It is not known whether the silver bromide or perhaps another substance is responsible for this kink in the sensitivity tail, and presumably most of the absorption beyond. The long-wavelength absorption tail varies considerably with the preparation of the emulsion, such as the heat treatment, and with the properties of the gelatine used.

Measurements of the spectral sensitivity of emulsions in the ultra-violet have been made by Vassy, Pinoir and Vassy (1945), Vassy, Schonne and Vassy (1947) and Vassy and Vassy (1947). Fresh emulsions rise continuously in sensitivity down to 2500 A.; aged emulsions develop a sensitivity peak, the position of which varies from 3800 to 4800 A., depending on the grain size and the density at which sensitivity is determined. The cause of the changes is so far unexplained.

4.6.4. *Adsorbed ions.* The absorption of light by silver halide crystals (sols) and the effect of various additions has been studied extensively (Fajans and Beckerath 1921, Fajans 1922, Fajans and Frankenburger 1923, Frankenburger 1923, Herzfeld 1923, Fajans and Steiner 1927, Steiner 1927). In these experiments the absorption was measured by the spectral sensitivity of photolysis which was followed by measuring both the amounts of silver and halogen produced as well as watching the visible discoloration. As was mentioned above (§ 4.3) silver halide crystals will adsorb various ions present in the surrounding liquid. Silver ions, in particular, cause an extension of the absorption tail into the red, whereas the addition of bromide ions produces no changes. It has been argued (Herzfeld 1923) that an electrostatic distortion of the surface is responsible for the change in absorption, but on this simple picture the addition of bromide ions should cause a similar extension of the absorption tail. The addition of OH ions also causes an extension of the absorption tail towards longer wavelengths. It is interesting to note that this extension, similar to that produced by Ag ions, disappears on continued exposure, thus indicating that the electron is actually released from the OH ions themselves which are thereby destroyed. Other ions behaving like OH, but of less practical importance are: phenylpropionic acid, SO_2, and Tl.

4.6.5. *Colloidal particles.* Another extension of the absorption tail of silver halides is produced by the presence of colloidal particles of silver or silver sulphide (Stasiw and Teltow 1941 a, b). The case of the silver particles is of particular interest since they occur as a direct consequence of exposure to light. It is apparent that these silver particles act in two completely different ways. In one the electron is released by a process of photoelectric emission from the silver particle into the conduction band of the parent crystal; as a consequence the silver particle itself is destroyed. The spectral sensitivity curve of this effect, which in the case of photographic materials is known as the Herschel effect (§ 2.2.8), is identical with the absorption band of metallic silver particles measured in large single crystals (Debot 1942 b, see Berg 1947).

The other process which has been overlooked for a long time in the literature consists of a direct transfer of the electron from the bromine ions into a silver speck near to them. This process also extends the spectral sensitivity although

the actual curve has not been measured; it is thought to be distinguished from the Herschel effect by the fact that halogen atoms are also formed (Frankenburger 1923), whereas with the Herschel effect the halogen ions are not involved in any way.

The change in absorption, as indicated by a change in spectral sensitivity brought about by the presence of silver or sulphide particles in or on photographic grains, has not been measured quantitatively and the process by which absorption is effected is not clear. The mechanism suggested below for dye sensitization may be effective. The spectral sensitivity for printing-out is relatively higher at long wavelengths than that for latent image formation for the same emulsion (Eggert and Kleinschrod 1940); this may be accounted for by the sensitivity for long wavelengths resulting from the increasing accumulation of silver with increasing exposure.

4.6.6. *Dye sensitization.* The natural sensitivity of silver halide gelatine emulsions is limited to the ultra-violet and blue region of the spectrum in which the silver halide crystals themselves absorb. The sensitivity of emulsions can be extended towards longer wavelengths as far as the near infra-red (1·1–1·3 μ) by the addition of certain sensitizing dyes to the emulsion, either during the making or by bathing of the ready-made material.

A sensitizing dye, in order to be technically useful, must fulfil a number of requirements. It must absorb light of the required wavelength; it must be adsorbed to the silver halide grains; it must not have serious chemical desensitizing or fogging action. Apart from these obvious requirements, a large amount of material has been collected which will allow the classification of these dyes. The subject is so wide that reference should be made to articles specially written on it (Mees 1942 a, chapters 23, 24, 25).

The colour of organic materials has been demonstrated to be intimately linked with their structure (see Brooker 1942). The colour appears to be associated with the occurrence of chains of conjugated double bonds, i.e. alternations of single and double bonds between carbon atoms, finishing up at both ends with suitable terminal atoms, usually oxygen or nitrogen, which can change their covalent linkage to carbon readily from one to two. An electron can be transferred from one nucleus to the other by a mere switching round of the order of the single and double bonds, and in some instances may be considered to be so transferred by the action of light. It has been pointed out (see Sheppard 1942) that for this process to occur efficiently the dye must be planar, a configuration which lends itself to the formation of aggregates (nematic state), in particular when the dye is adsorbed to a surface. The arrangement of the dye molecules on a silver halide surface can then be imagined to be similar to a stack of cards. In this state of aggregation, new specific absorption bands occur, as a rule extending the absorption towards longer wavelengths. Efficient sensitization is possible, however, in the absence of aggregates of dye molecules. On absorption of light, the electron in the dye moves along the plane; a similar induced electron displacement in the silver halide surface on which the dye is adsorbed has been made responsible for the functioning of dye sensitization.

West and Carroll (1947 a) have shown that an efficient dye-sensitizer increases the photoconductivity of a photographic emulsion, and that the spectral sensitivity

curves of photoconduction and photographic sensitivity are closely parallel. They have also accounted for the action of the so-called supersensitizers (1947 b). These bodies have the property of increasing the efficiency of dye sensitization (Bloch and Renwick 1920, Mees 1937) without themselves absorbing light or necessarily being sensitizers. Photoconduction and photographic sensitivity are increased in a very similar fashion by the presence of supersensitizers. These bodies are thought to be dispersed mono-molecularly in the superficial dye layer and to facilitate the transfer of electron energy from the excited dye molecules to the silver halide crystal. The authors have also shown that desensitizing dyes do not appreciably affect photoconductivity under conditions where photographic sensitivity is practically completely repressed.

The colour of dyes is slightly affected by their surroundings and is therefore not identical, say, for a dye crystal and a layer of dye adsorbed to silver halide. The spectral sensitivity of the latter system has been shown to run parallel with its absorption (Leermakers, Carroll and Staud 1937 a, Leermakers 1937).

The mechanism of latent image formation, apart from the primary act of absorption of light, appears to be identical in the dye sensitized and the natural spectral sensitivity regions. Thus the shape of the reciprocity failure curves for different wavelengths is identical for any one material (Webb 1933 b); the curves would no doubt be completely identical if it were possible to plot them not in terms of (visual) intensity but of numbers of quanta absorbed. Similarly the distribution of the latent image as between the surface and the interior of the grains (to be discussed below, § 5.4.2), is substantially identical for all wavelengths, provided the same density is produced by exposures of the same duration. All these results indicate that there is no essential difference in the process of latent image formation whether the original absorption occurs in the crystal or in the dye molecules. This fact is known as the van Krefeld law (van Krefeld 1933, 1934) and has been re-formulated by Webb (1936).

In experiments on printing-out it has been shown that one dye molecule can be responsible for the successive production of as many as fifty atoms of silver (Lescynski 1926/7 a, Tollert 1929). This means either that the dye molecule regains an electron after it has passed its own electron over to the silver halide crystal, or that it does not actually lose an electron during the absorption act but passes its excitation on to the nearest bromide ion in the grain. The energy deficit in the latter case would have to be supplied by thermal energy at the moment of energy transfer. Thus one would expect that supply of thermal energy is an essential condition for dye sensitization to function and that it would break down at sufficiently low temperatures. This has been found in certain instances (Sheppard, Wightman and Quirk 1934, Ungar 1937) but not in others (Webb and Evans 1938).

The mechanism of dye sensitization still awaits elucidation. It has been suggested (Gurney and Mott 1938) that the energy levels of the dye are so situated that the excited level is just above the lowest conduction levels of the silver halide. The photo-electron could then be transferred easily from the dye to the crystal. If this mechanism held, the electron would be lost to the dye and would have to be replaced. Recently, however, Coulson has pointed out (in a private communication) that the energy levels of the dye must be situated at least some

six or seven electron volts below zero, corresponding to the ionization energy of most organic molecules. Thus the energy transfer must be by a different mechanism altogether.

Mott has shown (1946 and private communication) that such energy transfer is possible in spite of the large energy deficit which is apparently involved (see also Franck and Teller 1938). According to his picture, the bromide ions themselves, on the surface of the crystal near the dye, will have to have a very small absorption in the spectral region in which sensitization is to occur. They will then be able to "resonate", i.e. a mechanical transfer of energy becomes possible; the probability of energy transfer by collision is very high, even if the absorption of the bromide ion in that spectral region is only about 10^{-6} of the normal absorption in the blue. The point is that it takes some 10^{-8} second for the dye to re-radiate the quantum of light absorbed, during which time there may be many collisions between the dye molecule and its surroundings.

The absorption of light of long wavelengths by the bromide ions near the surface may be thought to be brought about in various ways. Thermal energy may be sufficient to distort the structure for the short time involved; alternatively, there may be some permanent distortions of the grain surface brought about by impurities which cause the absorption at long wavelengths. In this connection, the kink in the absorption tail demonstrated for photographic emulsions (Eggert and Biltz 1940, Eggert and Kleinschrod 1940) mentioned above acquires a special significance.

Finally, the possibility must be considered that the state of excitation brought about in a dye molecule by absorption of light could be transferred to neighbouring dye molecules long before there is a chance of re-radiation, in particular, since the dye molecules are likely to be arranged parallel to one another and are thus very closely coupled (Scheibe 1937). One could imagine, therefore, that the state of excitation will travel around the surface of a dye-sensitized crystal until it meets a bromide ion ready to take over the energy, thereby activating the ion so that it emits an electron into the conduction band of the silver halide crystal. From then onwards the process of latent image formation will necessarily be identical with that occasioned by absorption of light by a bromide ion. One would expect that both silver and atomic bromine occur as the result of exposure to light, regardless of whether absorption is by the crystal itself or by a dye molecule on its surface.

4.7. *Fluorescence*

At sufficiently low temperatures photographic emulsions will fluoresce, the maximum of the emitted band being at 540 mμ (Meidinger 1940, 1943). The temperature at which fluorescence begins appears to depend on the grain size, being lower the smaller the grain. For a coarse-grain emulsion, the quantum efficiency of this process is of the order of unity. The process accounts for most of the loss in sensitivity of photographic materials at very low temperatures, assuming that fluorescence is due to recombination of electrons with halogen atoms. From the colour of the light, however, it is more likely that the fluorescence is due to trapping of electrons (see §7.2). No similar experiments have been reported with single crystals of silver halides.

§ 5. LATENT IMAGE FORMATION

5.1. *Basic Mechanism*

The theory by Gurney and Mott (1938) accounts for the formation of the photographic latent image by considering the two processes already mentioned: photoconductivity and ionic conductivity. By a combination of these two processes it is easy to visualize a mechanism whereby light absorbed all over the crystal produces one or a few small specks of metallic silver, situated in the main on the surface of the grains where the developing solution can reach them. The link combining these two processes was suggested to lie in the discovery that photographic emulsions are relatively insensitive to light unless certain impurities are present during emulsion making. The impurities were shown to be capable of forming specks of silver sulphide (Sheppard 1925), largely on the surfaces of the grains; these have become known as "sensitivity specks" (see § 5.3). For a long time the silver sulphide specks had been regarded as nucleating centres on which the photolytically-produced silver collects; they are now assumed to act by trapping the electrons liberated in the crystal by absorption of light. The centres thus become negatively charged and attract the mobile interstitial ions which are thereby discharged. In this way a speck of metallic silver collects on the silver sulphide speck as shown diagrammatically in figure 5.

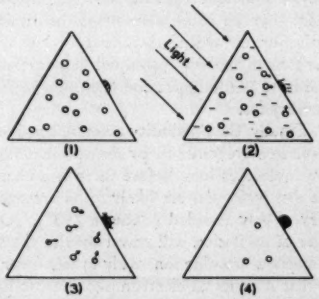


Figure 5. Diagram of latent image formation according to Gurney and Mott. The triangles represent a silver halide grain, ○ = interstitial ions; − = photo-electrons; shaded areas = sensitivity specks; solid area = latent image.

It is thus assumed that in a sensitivity speck electrons can occupy states of energy lower than those corresponding to the conduction levels of the silver halide. There is so far only indirect evidence supporting this assumption. Silver selenides and tellurides can also act as nucleating centres (Sheppard 1925), and it is probable that other impurities, as well as mechanical imperfections, are similarly capable of trapping electrons. Experiences with large crystals (§ 4.5.1) indeed suggest that it is hardly necessary to assume the existence of separate chemically identifiable traps, since effective trapping, coupled with the production of colloidal silver particles, occurs even in clean crystals, provided that physical imperfections exist (Hilsch and Pohl 1930). Annealing of the crystal causes these physical traps to disappear.

Electron traps of the latter kind appear to come into action when the more efficient silver sulphide specks are either overloaded or do not exist. This would seem to be obvious from the fact that latent image silver is also formed inside the grains where a normal developer does not reach, and where silver sulphide specks do not seem, and judging from the conditions of manufacture are not likely, to occur (§ 5.4).

Very plausible descriptions of many photographic facts can be given in terms of this fundamental mechanism. Furthermore, various predictions could be made on the basis of the new mechanism and were eventually tested and confirmed, to mention only the absence of reciprocity failure at sufficiently low temperatures (§ 7.2) and for sufficiently short exposure times at room temperature (§ 7.1). Those phenomena which cannot be so explained appear to be of a kind requiring subsidiary assumptions rather than changes in the fundamental mechanism; there has been in fact no serious criticism of the present theory since it has been put forward. Detailed aspects of the way in which the mechanism accounts for the more important photographic facts will be found in the body of the Report below (§ 7).

5.2. *Size of Latent Image Specks*

It is commonly assumed that in order to be able to induce development of a grain a latent image silver speck must exceed a certain critical size. Evidence for the fact that the latent image speck is formed in at least three stages will be mentioned below (§ 7.1). It is impossible, however, to determine the sizes of the specks directly because they are so small. The most useful estimates are based on extrapolations from the amounts of silver produced for very heavy exposures, sufficient to cause printing-out. For photolysis a quantum efficiency of unity is assumed, and has in fact been found to be valid under favourable conditions, i.e. in the presence of halogen acceptors, in the classical investigations by Eggert and Noddak and their school (Eggert and Noddak 1921, 1923, 1924, 1925, 1927 a, 1929, Kieser 1928/9, Mutter 1928/9). These estimates can only give average figures for the various grains, and because of the extrapolation involved the data merely constitute a limiting case. The shape of the toe of the characteristic curve has been used as the basis of other calculations (Svedberg 1920/1, Svedberg and Anderson 1921, Silberstein and Trivelli 1938, Brentano and Baxter 1940, Silberstein 1941, 1945, Webb 1941, Trivelli 1946 b), containing in many instances tacit and unwarranted assumptions on the mechanism of latent image formation and growth; it has been conclusively demonstrated that no unique conclusions can be drawn in this way regarding the number of quanta absorbed, and thus the likely number of silver atoms necessary to form a latent image (Webb 1941).

Reinders and Hamburger (1932, 1933) have attempted to determine the size of the latent image speck by sputtering metallic silver on to glass and "developing" it by depositing silver from a "physical developer" plating bath. By counting the number of developed particles and correlating it with the calculated number of chance aggregates of different sizes, they demonstrated that a development nucleus must contain at least four silver atoms. The evidence has been discussed recently (Berg 1945); the most doubtful assumption is that of chance aggregation, since the atomic silver is likely to be mobile for some time after deposition, and larger particles may thus be formed.

Sheppard (1931) has reviewed the question comprehensively and concludes that at the threshold of the characteristic curve one or at most quite a few quanta are absorbed by a grain. On the average, at a density of one, where perhaps one-third of the grains in an emulsion has been made developable, a few hundred

and sometimes several thousand quanta have to be absorbed by a grain. This does not necessarily mean that the minimum size of any one latent image speck itself varies within these limits. The most sensitive grains may already, before exposure, carry on their surface almost all the silver required to cause developability; on the other hand, the low sensitivity of some grains may be only apparent since they may be situated at the bottom of the emulsion layer. Furthermore, the amount of silver formed on any one grain may be distributed over several competing electron traps, thus producing several latent image specks on a grain. One would conclude, however, that it is necessary to account for the possibility that an appreciable number of grains in an emulsion can be made developable on formation of only a few silver atoms.

In contradiction to this conclusion, it has been claimed that it is possible "to see" under the ultra-microscope the colloidal silver particles produced on exposure (Savostjanova 1935). This would require a diameter of the order of 1000 A. To resolve this apparent contradiction, it has been suggested that the particles are visible by a fluorescence process (Berg 1938 b). Electrons are raised (in energy) from the silver to the silver halide conduction levels and drop back with the emission of light. This question has not been finally settled.

Another source of information on the number of quanta required to form a latent image on a grain, is available from investigations of the intermittency effect (§ 2.2.4). The photographic density produced on exposure to intermittent light has been found to vary with the frequency of the flashes used. There is, however, a critical frequency above which the density remains constant (Webb 1933 a). It has been argued (Silberstein and Webb 1934) that this critical frequency must be governed by the number of quanta received by a grain during the whole of the exposure, the light itself being received intermittently becasue of its quantum nature. Beyond the critical frequency, interrupting the light does not produce any difference in the degree of intermittence of the arrival of quanta. In this way, the number of quanta necessary is estimated to be of the order of 100 for normal photographic emulsions.

Again, it is not necessary to assume that this actually corresponds to the number of silver atoms making up a latent image speck, for reasons already given.

Finally, the number of silver atoms necessary for developability can be estimated (James and Kornfeld 1942) from work on X-ray exposures by an argument involving the number of ionizations per grain produced by a single X-ray quantum which cannot be discussed here in detail. Again, it appears that approximately 30 silver atoms must be formed to produce a latent image speck.

5.3. *Formation and Size of Sensitivity Specks*

The recognition of the nature and the investigation of the mode of formation of the sensitivity specks was one of the greatest advances connected both with latent image research and with the art of emulsion making (Sheppard 1925, Sheppard and Hudson 1927/8). Sheppard and his collaborators have isolated the sensitizing substances from the various by-products in gelatine manufacture. The amounts involved were so small that in the absence of any prior knowledge as to the nature of the sensitizers, direct chemical methods were inadequate to test for the presence of the substances. The procedure adopted was to follow the

various stages by emulsion experiments using photographically inert gelatines to suspend the grains, adding to the emulsion extracts of the various gelatine by-products and thus checking the presence of sensitizers by the increase in speed of the emulsion. It was found that photographically useful gelatines contain isothiocyanates or other labile sulphur compounds with the groups $S{=}C\langle$. These can form addition complexes with silver halides, which decompose on heat treatment to form silver sulphide.

The silver sulphide so formed is thought to be situated on the grain surfaces in the form of one or a few sensitivity specks. Evidence on the physical disposition of the sensitizer is scanty; when amounts much in excess of that necessary for optimal sensitization are used, actual specks of silver sulphide can be seen on the surfaces of the grain (Sheppard and Hudson 1927, Sheppard, Trivelli and Wightman 1927). The reaction leading to the formation of silver sulphide bears features of autocatalysis (see e.g. Carroll and Hubbard 1931 a), which also suggests that any silver sulphide formed will occur in the form of specks.

There are other good reasons for thinking that the sensitivity specks so formed are largely situated on the grain surface. During the stages of growth of the emulsion grains both the halide ion concentration and the pH are such as to impede the formation of silver sulphide (Carroll and Hubbard 1931 a). The final increase in sensitivity of photographic emulsions is brought about by continued heat treatment after washing, i.e. after the surplus halide ions which make grain growth possible have been removed. It is at this stage that the adsorbed complexes are thought to break down to form silver sulphide, the more completely the longer the heat treatment.

When the emulsion grains are grown in an inert gelatine, the speed can be increased by adding gelatine containing sulphur sensitizers and digesting; the finished emulsion may be as fast as one in which the grains are both grown and digested in the same "fast" gelatine.

Another indication that at least the silver sulphide fog specks are situated on the surfaces of the grains is provided by the Waterhouse reversal. The addition of labile sulphur compounds to the developer causes the image to be developed as a positive. This is thought to be brought about as follows (Rawling 1926). In a developer the conditions are favourable towards a breakdown to silver sulphide of the silver halide-thiocarbamide addition complex, provided that the bromide ion concentration is sufficiently low. In those areas which have not received an exposure, the conditions remain favourable for the formation of silver sulphide specks which grow until the specks cause the grains to be developed. In the exposed areas, the developing grains produce a concentration of bromide ions sufficient to repress the formation of silver sulphide, so that the final density is considerably lower than that in the unexposed areas.

It is interesting to consider the amount of silver sulphide present in an emulsion when optimum sensitivity has been achieved. Sheppard (1931) has reviewed this question and states that an average figure for the amount of silver sulphide is 5×10^{-4} gram per gram of silver bromide. This figure should perhaps be regarded as an upper limit since it is possible that not all the silver sulphide measured actually plays a part in the sensitization process or is even adsorbed on the grain surface. With a normal emulsion this amount of silver

sulphide works out to be sufficient to cover all the grains in an emulsion with a monomolecular layer or alternatively to produce one or a few sensitivity specks which are very large compared with the amount of latent image silver which on exposure collects on them.

If it is assumed, however, that the silver sulphide molecules are distributed at random over the surfaces of the grain, and further, that only the largest aggregates so formed will act as sensitivity specks, then it can be shown that the size of the sensitivity speck is approximately 10 molecules of silver sulphide and thus of the same order of magnitude as the latent image speck which is going to form on the sensitivity speck (Berg 1943 a).

As digestion time is increased beyond the optimum two effects are observed which are linked, yet distinct. First the sensitivity of the emulsion drops; second the fog value rises, i.e. there is an increasing number of grains which develop without exposure (Carroll and Hubbard 1931 a). The drop in sensitivity may be due to the formation of more than one sensitivity speck on an increasing number of grains, the specks competing with one another for the latent image silver (Sheppard, Trivelli and Loveland 1925, Sheppard 1928). The rise in fog may be due to the formation of oversized specks of silver sulphide which can by themselves induce development. We must thus conclude that a slight increase in the size of a silver sulphide speck will transform an optimum speck into a "fog" speck. In order to understand this it is necessary to assume that the electronic energy levels in a speck deepen rapidly with increasing size. According to the original conceptions of Gurney and Mott (1938) development occurs if the levels of latent image or fog specks are sufficiently low, while those of sensitivity specks of optimum size must lie above the latent image levels but be otherwise as low as possible.

It would appear that this picture is difficult to understand if a sensitivity speck consists of a very large number of molecules, whereas it becomes quite natural if the number of molecules involved is small.

5.4. *Disposition of the Latent Image*

The disposition of the latent image comprises its "dispersity" and its "distribution", defined with respect to single grains. By a highly dispersed latent image we mean the occurrence of a relatively large number of small specks on a grain; by the distribution we mean the topography of the latent image silver with respect to the surface and interior of the grain. In interpreting experiments carried out to investigate these properties of the latent image, it is as a rule assumed that the characteristic curve of the material is representative for the behaviour of a single grain. The diversity of grain sizes in an emulsion layer and the position of the various grains in the depth of the coated layer operate to make this premise invalid in certain instances; it is always essential to design the experiments so as to include a check on this point. The difficulty is increased by the fact that even photographic grains of the same size confined to a single layer differ in sensitivity, no doubt because of the different chemical sensitization they have received (Toy 1921). For this reason, considerations on latent image dispersity and distribution are usually confined to low densities where these difficulties are lessened.

5.4.1. *Growth and dispersity of latent image specks.* The growth of latent image specks in the course of exposure is followed most easily by plotting the density (number of developed grains) against the time of exposure (Burton and Berg 1946) rather than against the log (exposure) as in the characteristic curve (§ 2.1). The shape of this "*D–t*" curve can be interpreted very much like a chemical reaction curve in which the amount of reaction product is plotted against the time of reaction. For photographic materials the reaction curve exhibits a threshold and a toe; in other words, the addition of two exposures will produce a higher density than the sum of the densities due to each of the exposures applied separately. We can infer, therefore, that in the early stages of exposure a product accumulates, which later on is completed to form the latent image. The product formed in the first stage of this process has been termed the latent sub-image (not the "sub-latent" image!) indicating that a stable product is formed which, however, does not cause a photographic grain to be developable. The reality of this sub-image is indicated by the fact that it can be "developed" in various ways. First of all there are definite indications that with increasing degree of development a sub-image for a weaker developer may act as a latent image for a stronger developer; similarly a sub-image for a short time of development may become a latent image for longer times of development (Berg 1946). Secondly, it is possible to build up a sub-image to a full latent image by a carefully controlled post-exposure at very low intensities of light. For these low intensities those photographic grains which do not carry a sub-image speck are very insensitive (low-intensity reciprocity failure); the sub-image grains, however, utilize the low-intensity light with full efficiency.

The shape of the density–time curve (see figure 6), and the existence of the sub-image so demonstrated, imply the existence of hypersensitized grains at the end of an exposure. It is in these hypersensitized grains that the sub-image is built up to a full image by the low-intensity post-exposure. If it is assumed that all hypersensitized grains are made developable by the post-exposure, the *D–t* curve for a post-exposed material (*t* still being the time of the first exposure) represents a plot for *all* the sub-image produced during the first exposure. This includes the sub-image which has already been built up to the full image size by the first exposure. The curve goes straight through the origin and has its highest slope there (curve 2 of figure 6), facts which indicate that the sub-image is not subject to a process of gradual growth but is formed instantaneously. Under these conditions the density increase for a small interval of time during the first exposure does not depend on the previous exposure history but only on the actual exposure received during that time (leaving out, of course, the post-exposure, which is here regarded as a method of development). The condition for the formation of a

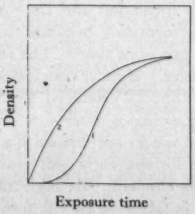


Figure 6. Diagrammatic density–exposure time curve for a photographic material. Curve 1 is that obtained under normal conditions, 2 that for a material post-exposed at low intensities or originally exposed at very low intensities without post-exposure.

sub-image speck appears to be that a certain minimum of quanta is absorbed by the grain in a certain maximum short interval of time; in order to form a sub-image a grain must receive a "burst" of quanta (Berg 1943 b). Sub-image formation may thus be considered to be a chance process.

At low intensities of light the probability of a burst being received by a grain decreases; this is considered to be the basic cause of low-intensity reciprocity failure. It is a moot point whether the instantaneous formation of the sub-image means that no silver speck is formed until a certain concentration of electrons is set up in the grain (Gurney and Mott 1938), or that a silver speck is actually formed as soon as any electrons are liberated and that the small speck so produced evaporates very rapidly (Webb and Evans 1940 b).

At sufficiently low intensities of light the density–time curve (without post-exposure) is shown to pass straight through the origin where its slope is highest (Burton and Berg 1946). The shape of the low-intensity curve is easily understood if it is assumed that the inefficiency of latent image formation at low intensities is confined to the sub-image stage, and that the process of building up the sub-image to the full image is fully efficient as is suggested by the post-exposure experiments. On this assumption, the time elapsing before a sub-image speck is formed occupies on the average a relatively large part of the total exposure; the completion of the speck to full image size a relatively small part. Thus at low intensities the sub-image and full image specks are formed almost instantaneously and the density–time curves for both are practically identical.

These results give an indication of the ratio of sub- to full image grains to be expected after an exposure carried out at different intensities of light. At very low intensities sub-image formation is relatively quickly followed by full image formation and the ratio of the two is low. At higher intensities sub-image formation becomes more efficient and takes up a smaller fraction of the total exposure time. The ratio of sub- to full image will thus go up with the intensity. Similar conclusions may be drawn about the average size of the latent image specks produced for different intensities of light. Sub-image formation being a chance process, at low intensities of light sub-image and thus full latent image specks are formed at a fairly even rate throughout the course of exposure; those formed early will grow to a large size, silver formation for them being highly efficient. At high intensities, on the other hand, most of the exposure time is taken up by the formation of the sub-image and, in consequence, most latent image specks will be completed towards the *end* of the exposure. There is then no opportunity for any of the latent image specks to grow to excessive size: latent image specks formed at high intensities will on the average be smaller than those formed at low intensities of light (de Langhe 1936).

These considerations are borne out by various experiments. For example, the density produced by a low-intensity exposure develops more rapidly than one produced for high intensities of light (Cabannes 1929, Hoffmann 1935, Strock 1938, Berg 1946, de Vaucouleur and Viollet 1947), a phenomenon which in its more general form is probably responsible for the Kron effect (§ 2.2.5). The destruction of the latent image by a red-light post-exposure (Herschel effect: see §2.2.8) is much more effective when the first exposure is to a high-intensity

of light than to a low, indicating that the latent image specks produced at low intensities are more stable (Lescynski 1926/7 b).

Similar results have been reported by Meiklyar (1941) on large crystals of silver chloride. Judging from the spectral distribution of light scattered by photolytic silver, he concludes that the size of the silver specks decreases with increasing intensity (and, incidentally, increases with increasing exposure time).

No definite conclusions have been reached regarding the relative dimensions of sub-image and full image specks. Webb and Evans (1940 b), who were the first to suggest clearly the existence of sub-image specks, have deduced from the exposures necessary that they are approximately half the size of the full image specks. This result has not been accurately confirmed (Burton and Berg 1946), but generally speaking, a stable sub-image speck may be imagined to contain between one quarter and half the number of silver atoms making up a full image speck.

Burton (1946) has shown that the sub-image fades on keeping, the smaller specks disappearing altogether, the larger ones growing into full latent-image specks, presumably at the expense of the smaller specks. He has obtained similar results for latent images situated on the surface or in the interior (see § 5.4.2) of the grains (Burton 1948).

The realization of the existence of the sub-image has become of some practical importance where it is necessary to obtain the utmost photographic speed. When the problem is one of producing high speed towards an exposure at low intensities, a uniform pre-exposure at reasonably high intensities (but not too high, in order to avoid the desensitization due to internal silver, § 5.4.2) should be used. This provides an amount of sub-image which will register the low-intensity light with full efficiency (hypersensitization, § 2.2.7). Alternatively, when the speed is required for an exposure at high intensities and short exposure times, it is possible to build up the sub-image to a full image by a carefully controlled post-exposure at low intensities (latent image intensification, § 2.2.6) (Moore 1941). In the ideal case no additional fog will be produced in either instance. In practice this ideal will not be achieved. The main reason is that the sub-image is not absolutely stable but subject to fading with time (Burton and Berg 1946, Burton 1946, 1948); thus there is a limit to the exposure time that can usefully be applied for the second exposure and the intensity cannot be reduced sufficiently to ensure that no sub-image, and thus latent image fog, will be produced during a low-intensity post-exposure. The speed increase obtainable also depends on other properties of the emulsion which influence the formation and stability of the sub-image.

5.4.2. *Distribution of the latent image.* It has been realized for a long time that latent silver is formed during exposure not only on the surface of the grains but also in their interior (Kögelmann 1894, Lüppo-Cramer 1921, 1927, Landau 1931, 1934, Kempf 1937, Berg, Marriage and Stevens 1941, Kornfeld 1941). Treatment in a normal developer will affect in the main those grains which carry a latent image speck on their surface. If, however, the emulsion contains some grains carrying internal latent image specks only, the silver halide solvents present in almost all commercial developers may allow a certain amount of the internal latent image to be revealed. The solubility of the grains varies with their composition and size, and the amount of solvent is different for various

developing solutions. In consequence, conflicting results are often obtained when the presence of the internal image is not taken into account.

In recent years a considerable amount of work has been done on the various factors which affect the distribution of the latent image. Techniques have been evolved which allow the surface and the internal latent image to be developed separately and together. A number of papers have been published on this subject in Belgium during the war and have become available to English readers but recently; these papers are the subject of a separate review (Berg 1947).

In principle, all techniques for investigating latent image distribution make use of the following interlocking system of processing solutions (Stevens 1942):

(i) A "surface developer" which contains no solvent for the silver halide and thus presumably reacts only to latent image specks present on the surfaces of the grains. This is also used to test for the destruction of the surface latent image by the bleach bath (ii).

(ii) A bleach bath which allows the surface latent image specks to be destroyed without having too drastic an effect on the internal specks.

(iii) Either a solvent bath followed by a non-solvent developer (Kempf 1937), or a developer containing a sufficient amount of silver halide solvent to lay bare and develop the internal latent image. Variations of internal development make use of iodizing solutions (Lüppo-Cramer 1927) which break up the grains and allow the developer to reach the internal silver; another method of internal development consists in physical development after fixation, which, like all methods of internal development, must be preceded by a bleach bath.

It has been suggested that it is necessary to define even more closely the distribution of latent image silver in the interior of the grains (Debot 1943: see Berg 1947). Within 10 lattice distances from the surface of a grain is said to be situated the sub-surface latent image and, at greater depths, the deep internal image. Investigations on the latter have been reported but recently, largely because it occurs only for exposures so heavy as to cause solarization for the surface image. Thus the deep internal image is not of great importance for the ordinary behaviour of photographic emulsions. It has been thought to be responsible for the occurrence of the second reversal of solarization which is obtained with ordinary developers, since it does not itself solarize (Debot 1941 a). It occurs simultaneously with the first traces of printing-out of a photographic emulsion, although it has been shown to behave differently from the print-out image. In the discussion below, no more specific mention need be made of the deep internal image; most of the properties of the internal image are probably shared by the sub-surface and deep internal images.

(a) *High-intensity reciprocity failure*

Probably the most interesting aspect of latent image distribution is its dependence on the intensity of the exposing light (Berg, Marriage and Stevens 1941, Debot 1942 a). At low intensities of light and long times of exposure, the latent image is situated preferentially on the surfaces of the grains; it is apparently under those conditions that the sensitivity specks are most effective. At high intensities of light and correspondingly short times of exposure, a much

larger amount of the latent image is formed in the interior of the grains. Whereas the surface image is subject to reciprocity failure at both high and low intensities, the internal image has so far been reported to exhibit only low-intensity failure. It has been demonstrated that the reciprocity failure curve for a material developed in a solution containing a certain amount of silver halide solvent is intermediate in shape between those for the surface and internal images (Berg, Marriage and Stevens 1941); contradictory results are thus liable to be obtained unless the constitution of the developer and the time of development are closely specified.

The change in distribution of the latent image with increasing intensity of light could have been forecast on the basis of the theory by Gurney and Mott, assuming that the efficient sensitivity specks are situated on the grain surfaces, whereas in the interior of the grains there are less efficient traps for the electrons. In the course of a very brief exposure to light, the relatively sluggish interstitial ions cannot move up to the charged sensitivity speck sufficiently quickly to be able to discharge the electrons at the rate at which they arrive at the speck. The rise in potential when a single electron is trapped by a centre is sufficient to ensure that any further electrons are repelled until the first electron is discharged by an interstitial ion. In consequence an electron cloud is built up in the grains. The repelled electrons are thus forced to utilize less efficient traps elsewhere in the grain or on its surface. In this way the latent image formed at high intensities is, generally speaking, in a more highly disperse condition and, in particular, is formed to a greater extent in the interior of the grain than at normal and low intensities.

(b) *Solarization and latent image distribution*

Another important property of the internal image has been mentioned already: the absence of solarization. Solarization, where it occurs, is confined almost entirely to the surface image; only on rare occasions is there a trace for the sub-surface image (Hautot, private communication). Solarization can be impeded by treating the emulsion by means of a halogen acceptor such as sodium nitrite before exposure (Abney 1882), or by means of a silver halide solvent after exposure and before development (Nafe and Jauncey 1940, Webb and Evans 1940 a). The total amount of photolytic silver as tested by chemical methods continuously increases in the solarization region of the characteristic curve (Arens 1931).

Earlier workers have favoured the "coagulation" theory of solarization, according to which the latent image is in a more compact form and thus less efficient when the specks are built up to a large size (Arens 1931, Meidinger 1932).

The most widely accepted mechanism of solarization assumes, however, that it is brought about by a re-halogenization of the surface image. It has been pointed out that the tendency for solarization increases in the order: silver chloride, bromide, iodide emulsions (Stevens, private communication); in this series, the halogen ions combine with gelatine and other halogen acceptors with increasing difficulty. Internal silver is formed in amounts equivalent to the halogen released. This concept is borne out by the fact that solarization is more pronounced for high intensities of light than for low (Webb and Evans 1940 a, Debot 1942 a), in parallel with the amount of internal image formed.

(c) *Latent image distribution and Herschel effect*

Recent work on the Herschel effect has shown that the bleaching of a latent image by a post-exposure to red light (to which the material must not itself be sensitive) is confined to the surface image (Falla 1941, Hautot and Falla 1942, see Berg 1947). If the internal image is affected at all, it reacts by an increase in density. As determined by chemical methods, the total amount of latent image silver present is not affected by a red-light post-exposure (Lescynski 1926/7 b, Tollert 1929, Trivelli 1929). The Herschel effect was therefore interpreted as a dispersion of the latent image. It now appears that the following changes in the disposition of the latent image silver may be brought about by a post-exposure to red light: a removal of the latent image from one part of the grain surface to another; a removal from the surface to the interior; a removal from one part of the interior to another; a removal from the interior to the surface. Of these possibilities, the fourth has recently been neatly demonstrated (Debot 1941 b: see Berg 1947). If a photographic material is heavily exposed to white light and then bleached to remove the surface image, it is found to exhibit a new sensitivity band in the red corresponding to the known absorption of colloidal silver particles in silver halides. Absorption of light in this band shifts the internal latent image silver to the surface of the grains. It seems reasonable to suppose that the normal Herschel effect constitutes a similar removal of the surface image to the interior or other parts of the surface of a grain, thus reducing the original size of the latent image speck.

(d) *Latent image distribution and Clayden effect*

Clayden (see Wood 1903) found that a pre-exposure of very high intensity and short duration desensitizes some photographic materials towards a subsequent exposure at lower intensity. The high-intensity exposure produces a certain amount of internal latent image silver which in the grains not carrying surface image specks acts by trapping electrons and thus competes with the surface sensitivity specks for the electrons produced by the subsequent exposure to light of lower intensity (Berg, Marriage and Stevens 1941). This explanation was substantiated by the finding that the desensitizing effect for the surface image is accompanied by a sensitizing action for the internal image. A similar effect occurs if a low temperature exposure is followed by one at room temperature, both of which can be at similar intensities (Berg 1939). This bears out the analogy between an exposure at room temperature and high intensity and one at low temperature to be considered below (§ 7.2).

(e) *Latent image distribution and Albert reversal*

If a photographic material is pre-exposed and the surface image bleached away, it is found that the exposed areas are selectively desensitized towards a second exposure (Clark 1923, 1924). If the first exposure is to an image and the second one uniform, the image is recorded as a positive. In general the bleach treatment causes some desensitization by itself, but on this is superimposed a further desensitization which is stronger the heavier the first exposure. The desensitization was first explained by assuming that the tendency for sensitivity to be destroyed by the bleach is enhanced when the speck is converted

into a latent image speck (Clark 1923, 1924). It has been shown, however, that most, if not all, of the selective desensitization is caused by the internal latent image produced during the first exposure (Stevens 1939). Thus the Albert reversal is more pronounced if the first exposure is very brief; it occurs quite as readily on emulsions which have been chemically denuded of sensitivity specks before the first exposure; it is confined to the surface image and does not occur if the material is physically developed after fixation. It is a peculiar fact that in photographic literature the Albert reversal and the selective loss in sensitivity when a pre-exposed material is subjected to a bleaching treatment are rarely discussed together, being apparently considered as distinct phenomena.

§6. DEVELOPMENT

The events comprising photographic development are difficult to summarize, largely because neither the chemical events involved nor the underlying mechanism have so far been elucidated in sufficient detail. In considering the process of reduction of silver halide to metallic silver and halogen ions, which constitutes development, two problems present themselves. What factor distinguishes the exposed from the unexposed silver grains so that only the former will be developed; and what are the special properties of the reducing agents which enable them to make this distinction? The evidence on the first factor, which by definition is the latent image, has already been discussed. The second question will be dealt with under the heading "Mechanism of Development".

In all considerations on chemical processes such as development, it is necessary to draw a sharp division between two distinct aspects, i.e. that which decides in which direction a possible reaction will proceed, and that which will merely influence the rate of this reaction. Development used to be regarded as a reversible reaction capable of reaching thermodynamic equilibrium; considerations of the energies involved were thus made to account for the distinction a developer makes between exposed and unexposed grains. Present workers have come to regard this distinction as a difference in the *rate* of development. Both points of view are of value, however, and will be considered below.

6.1. *Energy Balance*

Many of the views still held on the process of development were suggested by Sheppard and Mees as early as 1907. The process of development was shown to have many features of a reversible reaction, its kinetics following in a large measure the law of mass action. Thus the rate of development is often found to be proportional to the concentration of the developing agent, and bromide ions in small amounts decrease the rate of development since they are a reaction product. The activity of developers has often been classified by their "bromide potential", characterized by the amount of bromide ion necessary to stop development completely.

It is only in special cases, however, that development can be regarded as a strictly reversible reaction. This condition was fulfilled in experiments by Reinders (1934) in which the reducing system was an equilibrium of $Fe^{++} - Fe^{+++}$ ions (a buffered ferrous oxalate developer). He showed that a solution capable

of acting as a developer is characterized by a redox potential * lying between two well-defined limits. If the redox potential is too high the solution acts as an oxidizer and dissolves the latent image, if too low even the unexposed grains are developed. The development of partially developed grains can be continued by a solution with a higher redox potential than is necessary to initiate development of exposed grains. Reinders describes the process by a model in which the silver halide is assumed to be dissolved and the reaction to take place in the solution so that the reduced silver is precipitated on the latent image nuclei, in accordance with the supersaturation theory of development (Ostwald 1893, Abegg 1897, 1899). Although this model is a convenient one for the purpose of discussion, it does not follow that the actual process of development proceeds in this fashion; the thermodynamic considerations may apply whatever the mechanism by which development proceeds.

According to Reinders (1934) the critical value of the redox potential, above which there is no development, is 100 mv. In a recent paper, Vassy (1947) gives a value of 50 mv.; the reason for this difference is not clear.

It seems certain that the supersaturation theory does not apply to practical processes of development. The solubility of silver halides in technical developing solutions is much too low to allow the whole grain to be dissolved in the time a practical developer is allowed to act, which is a few minutes as against many hours in Reinders' experiments. The grains are reduced by the developer as individual units. The presence of a minute speck of silver on the surface of the grains does not of course affect the energies involved in the reaction; its only effect can be to alter the rate of reaction by its catalytic action. Considerations of the redox potential, therefore, do not allow us to draw any conclusions as to the mechanism involved.

This will perhaps be clearer if the redox potential is interpreted in terms of electron levels (figure 7). A developer may be considered as a solution supplying

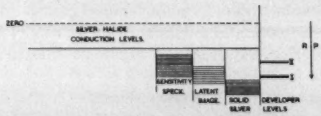


Figure 7. Electron levels in silver halide, sensitivity speck, latent image, solic silver, and for two developers I and II. RP denotes the redox potential.

electrons at a certain pressure; this can be denoted by an electron level which is higher the higher the reducing power of the solution, i.e. according to definition the lower the redox potential. If the reducing solution is brought into contact with a system which can itself be described as having an electron pressure, then electrons will flow from the higher to the lower level. Solid silver, the reduction

* The redox potential measures the reducing or oxidizing power of the system, when the reaction is assumed to proceed in a reversible way, as in a galvanic cell. It may be described as the tendency of a solution to set up an electrical double layer on a metal electrode, and is defined as

$$P = C/n \log (\mathrm{Ox}/\mathrm{Re}) + E_0,$$

where Ox is the concentration of the oxidized, Re that of the reducing ions, and E_0 the electrical potential usually measured against a standardized hydrogen electrode when Ox = Re; n is the difference in the charge of the ions between the reducing and the oxidized form.

product of silver halides, is a system that can be regarded as being capable of supplying electrons at a certain pressure. Thus, given time, reduction will occur if the developer level is above that of solid silver (I in figure 7). A small speck of silver (latent image) has higher levels than a large metallic electrode; a latent image speck would be dissolved, i.e. oxidized by a solution having levels at I. Although this fact was actually confirmed by Reinders, the grain must eventually be reduced by I, if the solution is allowed to act for an infinite length of time. The latent image speck acts as a catalyst and merely helps to overcome or decrease the activation energy. Once a solid speck of metal is formed, i.e. once the process of development is initiated, the reaction will proceed to completion. A solution with its electron levels at II will not tend to oxidize the latent image and will therefore at once initiate the reduction process, but it can do this only if the grain bears a latent image speck. Presumably a solution with even higher electron levels, so as to be able to supply electrons directly into the conduction levels of silver halide, is capable of reducing the grain without having to overcome an activation energy and thus constitutes an effective fogging agent. The difference in energy between the conduction levels and a solid silver electrode can thus be regarded as an activation energy which has to be overcome in order to initiate development in an unexposed grain. The difference between exposed and unexposed grains merely consists in the existence on their surfaces of a catalyst increasing the rate of reaction. Other surface conditions, however, also contribute to, and probably play a decisive part in, the activation energy involved.

For most practical developers the redox potential is considerably below that necessary to reduce unexposed silver halide. In consequence unprotected silver halide crystals are reduced immediately, regardless of any exposure they may have received (Sheppard 1929). In an emulsion the grains are protected by the surrounding gelatine as well as by their negative surface charge, as will be discussed below. In normal development the system is never given time to reach equilibrium and the reaction can rarely, if ever, be considered as a reversible one. This being so, the mechanism of development is of much greater significance than would otherwise be the case and studies on the rate of reaction are essential for the elucidation of this mechanism.

6.2. *Rate of Development*

6.2.1. *The density–time curve.* In recent years valuable information on the process of development has been gained by a more detailed study of the kinetics of the reaction. In this work curves of developed density or mass of silver were plotted against the time of development and have been interpreted in a way similar to the reaction curves of the chemist.

In general chemical procedure the interpretation of such reaction curves is well known: a threshold and toe indicate an autocatalytic reaction, which in the photographic case denotes that the presence of development products increases the rate of development. The absence of a toe indicates* a reaction of the first order, i.e. the rate is independent of the amount of reaction product and dependent only on the amount of substance still to be reduced. The

* But not necessarily proves.

interpretation of the reaction curves with photographic emulsions is not so simple, however, since the individual grains are normally reduced as discrete units. The development of the individual grains may be sub-divided into two stages: an induction period during which no visible development occurs, and the visible completion of the development process. The question arises, therefore, whether the shape of the density–time curve is due mainly to the progressive reduction of the individual grains as development proceeds, or whether it arises from the onset of development in one class of grains after another. In the former case the different grains in an emulsion pass through the various stages of development nearly at the same time, while in the latter case the curve merely indicates the distribution of the different lengths of the induction period for the different grains.

It had for a long time been accepted that the density–time curve actually indicates the distribution of induction periods. Meidinger (1935, 1936), using an emulsion with large grains in which development of individual grains could be followed, found that the induction period was long (up to 30 minutes) compared with the time of visible development which never exceeded 1 minute. This conclusion is contradicted by recent work due to James (1945 a), at least for emulsions of the slow, fine-grain type. He advances two arguments which indicate that the density–time curve can often be regarded as representative for an individual grain: plots of the amount of silver determined chemically against the density can be obtained in different ways. Thus, by fully developing a range of different exposures the variation can be brought about by varying effectively the number of grains so that no question of any partially developed grains arises in this case. Alternatively, and giving the same result, the variation in the number of grains can be obtained by using fully exposed and developed emulsion layers of different coating thickness. These data can be compared with the curve obtained for heavily exposed pieces of material developed for a range of times. The resulting curves are of different slope, the latter one coinciding with that worked out theoretically by Nutting (1913), Eggert and Küster (1936) for the case of grains increasing in size in the course of development. The second argument is similarly convincing. A pronounced toe can be obtained on a density–development–time curve by adding bromide ions to a hydroquinone developer. When, however, the bromide is added after development has started, but before an appreciable change in density has been obtained, practically no change is found in the density–time curve. This result is interpreted as indicating that the bromide ions affect the induction period for the grains, all of which start developing nearly simultaneously. If the exposure is sufficiently heavy the density–time curve can therefore be regarded as representing the course of reaction of a single grain.

This is no longer true for weaker exposures when the shape of the density–time curve is brought about both by the distribution in the length of the induction periods for the different grains and the course of development for each of the grains (James 1945 b). It has been found also (Berg, unpublished experiments) that latent image specks of different sizes as produced by exposures giving the same density on full development at different intensities of light, produce different rates of reaction. A "surface" developer was used in these experiments, and

the exposed grains were confined to the surface layer of the emulsion so as to avoid the delays entailed whilst the developer penetrates into the depth of the coated layer or of the individual grains. The interpretation of the induction period has been discussed recently by Bagdassarian (1944 a) who arrives at similar conclusions.

6.2.2. *Protective barriers.* The grains in a photographic emulsion are surrounded by two kinds of protective barriers, which in normal emulsions impart a negative charge to the grains. It is the presence of these barriers which must be considered to allow a developer to distinguish between exposed and unexposed grains. One of these barriers consists of surplus bromide ions adsorbed to the grain surfaces. A bromide barrier is present even if the silver halide crystals are prepared from stoichiometric proportions of silver and halide ions, since the hydration energy of the silver ions is larger than that for the bromide ions (see table 1, page 259). Normally however, a surplus of bromide ions is always present in ordinary emulsions. The second protective layer consists of the gelatine which is also strongly adsorbed to the grain surfaces. It is the adsorbed gelatine layer which determines the effective charge in electrophoretic experiments when the grains are held in suspension in a gelatine solution. Gelatine, being an amphoteric substance due to the presence in the molecule of both acid- and alkali-ionizable groups, provides a charge layer according to the pH of the surroundings. The grains are effectively negatively charged at high, and positively charged at low pH. They become neutral at the iso-electric point of the gelatine at which equal numbers of basic and acid groups are ionized. Since the iso-electric point of normal gelatine is well below the pH normally used in photographic emulsions, the sign of the charge on the gelatine is the same as that of the bromide ions. This means that both agents combine to surround the grains with a double-layer, the negative charges on the grain being neutralized by a floating population of positive ions in the surroundings of the grain. The presence of these charge barriers plays an important part at least in the kinetics of development.

It can be demonstrated, however, that it is not the charge of these barriers which is the only decisive factor in enabling a developer solution to distinguish between exposed and unexposed grains. We have mentioned the fact that silver halide grains prepared in the absence of gelatine and with a surplus of bromide ions so as to avoid coagulation will be reduced immediately by a normal developer regardless of whether they have been exposed or not (Sheppard 1929). It is possible to devise a developer capable of distinguishing between exposed and unexposed grains; this will have to be a very much weaker reducing agent or must contain a larger proportion of halide ions than is used in normal development. It is, therefore, the presence of the gelatine which is the decisive factor in the general case. James (1944) has shown that development proceeds still in the normal way at the iso-electric point of gelatine where its charge is absent. One must conclude, therefore, that the gelatine acts as a protective colloid (in the broadest meaning of the term) without being able to state any more clearly how this protective action with regard to developer solutions is brought about.

A series of papers has been published by James (1939, 1944, 1945 a, b, c, 1947) on charge barriers in photographic development, in particular on the effect of

the charge of the developing ion on the *D–t* curve (figure 8). Neutral developer molecules were shown to produce a *D–t* curve without an induction period; the induction period increases with increasing charge of the developer ion. On addition of bromide ions the induction period is much increased, as mentioned above, but this effect only occurs when the developing agent is a negative ion. The effect of adding bromide ions cannot be accounted for by mass action considerations, since it is absent with neutral developer molecules. The bromide barrier can be reduced by basic or cationic dyes such as phenosafranine or pina-cyanol, and also cyanines and thiocyanines in small amounts. Again, the effect only occurs with negatively charged developer ions. The photographic speed obtainable with an emulsion processed in various developers is also found to depend on the charge of the developing molecule (James 1945 c): when the material is developed to a constant gamma the neutral and singly charged developing molecules produce a considerably higher speed than doubly even or more highly charged ions. To avoid misunderstanding, this result must not be taken to mean that on full development the developing agents with ions of small charges will produce a higher speed than the highly charged ones: various additions such as that of sodium sulphite and potassium bromide also affect speed and gamma with technical developers.

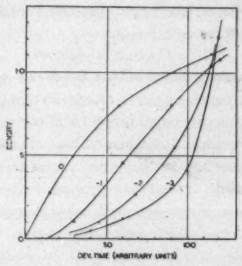


Figure 8. Curves of density against time of development for neutral developer molecules (0) and for singly, doubly, and triply (−1, −2, −3) charged ions. (*After James and Kornfeld*, 1942.)

On this basis James has attempted to explain various puzzling phenomena in development, of which only that of compound developers may be mentioned here. The density produced in a given time by a solution containing metol (Elon) and hydroquinone as developing agents is higher than the sum of the densities produced by similar solutions containing only one of these agents. This property of a compound developer is described as super-additivity, and is accounted for by James by suggesting that the singly charged metol will be able to penetrate the charge barrier more rapidly than the doubly charged hydroquinone. Once the charge barrier is broken down, i.e. the induction period has been overcome, the doubly charged hydroquinone is considered to develop more rapidly.

On the other hand, Levenson (1946) has found this explanation unsatisfactory as a result of experiments in which, *ceteris paribus*, metol was found to develop faster than hydroquinone even well beyond the induction period. Moreover, he finds that the metol-hydroquinone developer shows super-additivity when the rate of reduction of silver bromide is considered rather than the densities reached in a given time. This rate-super-additivity, which cannot be accounted for in terms of the induction period, is claimed to result from a regeneration of metol from its primary state of oxidation by reduction by hydroquinone,

The system *metol plus oxidized metol* acts as an electron carrier for the reduction of silver bromide by hydroquinone.

In a further paper Levenson (1947) studied a range of compound developers containing metol together with a reducing agent giving a polycharged ion. A consideration of the rates of reduction of silver bromide by the component developers, and the rate of oxidation of the reducing agents from the compound developers, led to the conclusion that the degree of rate-super-additivity correlates with the regeneration efficiency of the mechanism referred to above in the case of the metol-hydroquinone developer. Levenson (1947) shows that a number of dyes in the leuco form, including safranine-T and methylene-blue, are developing agents and suggests that the accelerating action of such dyes on hydroquinone developers may in fact be a further instance of super-additive development. After showing that gallic acid, resorcinol, and other reducing agents commonly regarded as being photographically inert can act as developers *via* safranine-T and methylene-blue systems, he suggests as a general definition that super-additivity in compound developers is manifested when the reduced form of one of the components can readily reduce exposed silver bromide in gelatine, and the other component can reduce the oxidized form of the first.

The presence of charge barriers also plays an important part in the considerations of the mechanism of development to be considered below.

6.3. *Mechanism of Development*

6.3.1. *The developed silver.* Most of the published work on development has but an indirect bearing on the mechanism of development. The most positive result is the finding that the developed silver occurs as a rule in the form of a tangled mass of ribbon-like filaments which can sometimes be seen in photomicrographs (Scheffer 1907), and are readily resolved by the electron-microscope (Hall and Schoen 1941). One gains the "overwhelming impression that the silver filament is formed by ejection from the silver bromide crystal" (Mees 1942 b). In layers of the thickness used for ordinary emulsions the developed silver occupies the space previously taken up by the silver halide grains, presumably because the gelatine forms a pocket which the thin filaments are not able to penetrate; in some cases the developed grain appears to be a perfect replica of the original one when viewed at low magnification. In thin layers the developed grains often occupy much more space than the original emulsion grains. These findings fit in well with the previous results mentioned above (§ 3.2) that visible development always begins at one or a few isolated spots on the surfaces of the grains.

6.3.2. *Electrolytic mechanism of development.* Gurney and Mott (1938) have suggested a mechanism of development which forecast the production of silver ribbons, and which was borne out by the electron microscope tests already mentioned. They regard the latent image as an electrode which accepts electrons from the developer and consequently attracts interstitial silver ions from the silver halide lattice, in analogy with the mechanism of latent image formation. In this way the original latent image is pushed out from the grain by the new silver forming underneath; thus a thread of silver is formed. The main objection to this mechanism has been that it does not account for the removal of the bromide

ions which cannot themselves move through the lattice (James and Kornfeld 1942); a space charge would thus be set up and would bring the process to a stop. Secondly, the rate of formation of interstitial ions is too slow to account for the rate of development (Berg 1943).

It must be considered, however, that the field strength near a latent image speck charged by a single electron is of the order of a million volt/cm. This is sufficient to disrupt the lattice by pulling out the nearest silver ion. The holes so formed are effectively positively charged and will be repelled by the charged silver speck; the migration of these holes through the lattice has been considered above (§ 4.4.2). The holes will accumulate on the parts of the crystal removed from the growing silver speck: in other words, there will accumulate on the surface a surplus of halide ions which are taken up by the solution surrounding the grain. This process is illustrated in figure 9. It amounts to a disintegration of the crystal, in which, however, all movement of silver ions occurs in the solid, the halide ions not being required to move through the lattice. This picture must, however, be regarded as tentative.

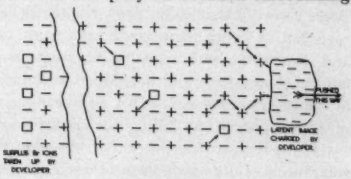


Figure 9. Electrolytic mechanism of development. (*After Berg*, 1943.)

It has been suggested (James *et al.* 1934) that in this and similar processes, both adsorbed silver ions and silver ions in solution might take part. Due to the big hydration energy of silver ions those in solution will probably not participate; no specific details are known as to the behaviour, in particular the mobility, of adsorbed ions.

Objections (James and Kornfeld 1942) have been raised to the suggestion that the latent image merely acts as an electrode accepting electrons from the developer; it is pointed out that adsorption phenomena play a part in development, developer ions or molecules being adsorbed on silver halides. This objection does not appear to rule out the mechanism completely, since in order to be able to transfer its electron, the developer molecule may only have to be adsorbed near to the latent image speck. Adsorption of developer molecules near the latent image is also a feature of another possible mechanism of development to be considered below.

6.3.3. *Development: an interface catalysed reaction*

James and Kornfeld (1942) have pointed out that development may not be as mysterious a process as is often made out: it falls into the general class of heterogeneous reactions, catalysed by the presence of one of the reaction products and occurring at the meeting place of the three components, viz. silver halide, metallic silver and developer. The silver ions in the silver halide are regarded as being "adsorbed" to the silver, this adsorption being the essential preliminary enabling the metallic silver to catalyse the reaction. Silver filaments are formed because the reaction is assumed to proceed along cracks and crevices of the silver halide. It has been pointed out (Berg 1943) that although this picture accounts for many facts in development, the explanation of the form of the developed

silver is not very satisfactory, since cracks and fissures would not be expected to be sufficiently regular. It was suggested that well-known observations on crystal growth may account for the shape of the developed silver. Adsorbed material in the surface layer of a crystal has been demonstrated to have a high lateral mobility (Volmer 1932, Berg 1938). We could imagine, therefore, that the newly formed atoms of silver could move along the growing thread and are fitted into the structure only at the fresh faces of the growing single crystal of silver the thread most likely represents. There is thus no need to assume movement of the triple interface where reaction occurs; this picture has been discussed more fully by James (1943).

On the basis of present knowledge, it appears to be impossible to decide upon a definite picture of the mechanism of development. The question is one of considerable importance both for its own sake and because the final answer is likely to influence considerably our concepts as to the nature of the latent image and its formation.

§ 7. PHOTOGRAPHIC PHENOMENA DISCUSSED IN TERMS OF THE GURNEY-MOTT THEORY

7.1. *Reciprocity Failure*

The explanation of the various aspects of reciprocity failure has been one of the most striking successes of the theory of latent image formation by Gurney and Mott (1938) and will therefore be considered in more detail here. The latent image, according to this theory, is built up by two distinct processes, one electronic, the other ionic. These processes have different time constants and temperature coefficients, the ionic process being more sluggish and more strongly dependent on temperature than the electronic. At different intensities of light and exposure times, the interplay between these processes varies; it is in this way that the failure of the reciprocity law and its change with temperature has been accounted for.

Most of the detailed knowledge on the breakdown of the reciprocity law has been obtained by Webb, but additional facts have been brought to light as a result of predictions based on the Gurney-Mott theory. According to this theory, the photographic process is less efficient at low intensities of light because the sensitivity will not be able to hold electrons for an indefinite length of time; a certain critical electron vapour pressure is necessary before a latent image speck can begin to grow. Webb and Evans (1940 b) proposed a slightly different version of this picture by suggesting that in its early stages of growth a latent image speck is unstable and will evaporate and be lost again unless further silver is supplied at a rate sufficient to build up the speck to a stable size.

At high intensities on the other hand, the photographic process becomes inefficient because of the sluggishness of the movement of the interstitial silver ions and the limited electric capacity of the sensitivity specks. The movement of electrons is very rapid, while the movement of silver ions takes more time, so that the charged sensitivity speck is not neutralized at once. When charged it will repel any further electrons; in this way a high concentration of electrons will be set up in the grain. This may lead to two possible consequences: either the condition for sub-image formation as set out above (§ 5.4.1) might become

fulfilled for shallow traps existing in the grain, or the electrons might cause a local breakdown of the lattice and thus form a silver speck. Which of these alternatives, if any, describes the mechanism correctly we do not know. We have already seen above (§ 5.4) that the latent image due to an exposure at high intensities appears to be formed in a relatively highly dispersed condition. High-intensity failure occurs mainly in the second stage of latent image formation, sub-image formation not being subject to high-intensity failure, as is demonstrated by the observation that high-intensity failure can be eliminated by a post-exposure at sufficiently low intensity (Burton and Berg 1946).

The results on reciprocity law failure can be used to estimate a lower limit to the number of silver atoms necessary to make up a latent image speck, by considering the number of distinct stages necessary to account for the different aspects of reciprocity failure. The sub-image is subject to low-intensity failure only; in order for this to be possible the sub-image must be formed in at least two stages and, if assumed to consist of silver, must contain at least two atoms. The completion of the latent image is seen to be subject to high-intensity failure; again it can be argued that at least two silver atoms must be involved in order to produce the effect. For an emulsion subject to both low- and high-intensity failure, the latent image speck should therefore consist of at least four atoms of silver.

At very high intensities of light a condition should be reached where in any one grain the electronic and ionic processes cannot be separated any further by shortening the time of exposure; the electronic process is practically instantaneous, its early stages being over before the ions have time to move appreciably. There should be a limiting time of exposure below which the process of latent image formation should not be affected by the time of exposure and thus not be subject to reciprocity failure. This was confirmed by experiment (Berg 1940): for a range of materials the reciprocity failure curve was found to bend over into the horizontal at an exposure time of about 4×10^{-5} seconds. The position of the bend-over point depends on the mobility of the silver ions and varies exponentially with temperature (see, however, Webb 1942).

The effect of temperature on reciprocity law failure (Webb 1935) confirms the general concepts of the Gurney-Mott theory, at least in a qualitative fashion. At reduced temperatures the ionic mobility drops and the rate of discharge of the sensitivity specks decreases: the electrons are therefore repelled from the sensitivity specks at a time of exposure for which at normal temperature the mobility of the ions is quite sufficient to discharge the electrons. Thus at reduced temperatures the whole high-intensity branch of the curve is shifted towards long times of exposure. For elevated temperatures the reverse process occurs and the curve shifts towards shorter times (see figure 2). The variation of reciprocity failure at low intensities is similarly affected. Here a small speck of silver becomes less stable at high temperatures, the probability of an electron being lost from a speck in a given interval of time depending exponentially on the temperature. Thus the low-intensity branch of the reciprocity failure curve is also shifted towards longer times of exposure with decreasing temperature. The temperature coefficients of reciprocity failure at high and low intensities

being different, one would expect some change with temperature of the shape of the curve; the most important effect, however, is a bodily shift of the whole curve towards longer time and lower intensities as the temperature is decreased, and towards shorter times as it is increased.

7.2. *Photographic Sensitivity at Low Temperature*

At sufficiently low temperatures the flat high intensity end of the reciprocity failure curve occupies the whole range of exposure times practicable (Berg and Mendelssohn 1938, Evans and Hirschlaff 1939); in other words reciprocity failure disappears entirely at temperatures as low as that of liquid air (90° K.). By exposing at these temperatures, part, at least, of the electronic process can be completely separated from the ionic one because of the drop of ionic mobility with temperature. In this way the electronic process can be studied separately.

Various factors collaborate to cause photographic sensitivity to be reduced for exposures at very low temperatures. The most important item is the fluorescence process discussed above (§ 4.7). Any further drop in sensitivity is presumably caused by the fact that at low temperatures one works effectively at very high intensities. In confirmation of this it was found (Berg, Marriage and Stevens 1941) that internal image formation largely predominates for a low temperature exposure, the gamma and speed being several times that for the surface image. In this case it seems probable that only a small proportion of the exposed grains containing internal latent image also bear surface latent image specks. The proportion of internal image at low temperature was, however, much higher than for the shortest exposure time at room temperature; it is thus evident that for a low-temperature exposure a process becomes effective which specially suppresses the formation of surface image.

The process of latent image formation is not yet stabilized, however, at 90° K. There is an additional decrease in contrast and maximum density at 20° K. (Berg and Mendelssohn 1938); photographic sensitivity still persists at 4° K. although no quantitative data are available (Berg 1939). From large crystal experiments it is known that photo-conductivity stops at 4° K. and one would expect photographic sensitivity to do the same; we do not know the processes occurring at these low temperatures (Lehfeldt 1935, Pohl 1937).

Experiments were done in order to investigate the nature of the traps in which the electrons are held at these low temperatures. It is clear that the electrons must be held trapped somewhere in the crystal because if they were free to move they would quickly recombine with the bromine atoms whence they came. The bromine atoms must also be trapped (§ 4.5.4). The electrons could conceivably be trapped either in the sensitivity specks themselves or in other traps which are sufficiently shallow to release the electrons on warming the emulsion. On warming, both the electrons in the shallow traps and the interstitial silver ions become free to move and the latent image is formed. Both kinds of traps appear to be operative. The effect of an exposure at low temperature—large enough to give an appreciable density on warming and development—consists of course, of trapped electrons; it can be bleached out very effectively by red light if the material is kept at low temperatures all the time. After warming and re-cooling this is no longer possible (Webb and Evans 1938, Berg 1939).

Exposure to red light thus appears to lift the electrons from the traps, when they will either be trapped afresh or will recombine. The fact that bleaching by red light is possible demonstrates that bromine atoms are still available for recombination at low temperatures. If the material is exposed to radiation containing a proportion of actinic as well as bleaching light, one would expect a state of equilibrium to be reached on continued exposure where equal numbers of electrons are liberated and lost in a unit of time, assuming that there is but one kind of trap. Actually a latent density post-exposed to mixed light first decreases to a minimum and then increases again on continued exposure, all at low temperature. This was interpreted to mean that the red light transfers electrons to deeper traps from which they cannot be liberated any more. Judging from the spectral sensitivity of the bleaching effect the deep traps must be between 1 and 2 ev. On the other hand, the shallow traps cannot be deeper than about 0·8 ev., otherwise the electrons would not become free at room temperature. Since the electrons are not freed from the deep traps on warming, one must assume that the deep traps are the sensitivity specks themselves, otherwise the electrons could not contribute to latent image formation (Berg 1939).

For low-temperature exposures photographic materials do not appear to be subject to solarization; in consequence a higher maximum density is sometimes reached in spite of the loss in contrast and speed (Berg 1939). Webb and Evans (1938) have demonstrated that the loss in speed must be largely due to the electrons being repelled from sensitivity specks which are charged by capture of an electron. They found that if the exposure is sub-divided, with periods of warming between the individual exposures at low temperature, speed and contrast approach that of a material exposed at room temperature.

The mechanism of latent image formation at low temperatures can thus be summed up as follows:

The exposure liberates electrons which move about in the conduction levels until they are trapped. There are deep traps of 1 to 2 ev. which constitute the sensitivity specks, and shallow traps of less than 0·8 ev. from which latter the electrons are released on warming. Frozen ions in interstitial positions then also become mobile and the latent image forms in the usual way. It seems essential to assume that bromine atoms also become frozen at low temperatures in spite of the fact that the movement of bromine atoms is thought to occur by an electronic replacement process. It also appears that the bromine atoms do not leave the crystal at low temperatures, otherwise no recombination could occur. There is so far no independent experimental evidence on this point.

7.3. *High Energy Quanta and Particles*

Practically the whole of the present report has been devoted to the effect of light on photographic emulsions, which is characterized by the fact that a single quantum of light has insufficient energy to make a grain developable. The photographic process, however, is not confined to light, and is applied to an increasing extent to the recording of high energy quanta as of x rays and gamma rays, or of particles such as electrons, protons, α-rays, and even mesons and fission fragments. With all these as a rule no more than a single quantum or particle is required to make a grain developable, and certain fast particles will

even cause whole string of grains (tracks) to be made developable. Of all these agencies x rays have been studied most extensively. Since a single quantum suffices to make a grain developable there is no reciprocity failure with x rays (Friedrich and Koch 1914), i.e. time- and intensity-scale characteristic curves can be used interchangeably. The characteristic curves for x rays are identical in shape for all wavelengths (Glocker and Traub 1921) except the very longest (Broili and Kiessig 1934) if referred to the number of quanta absorbed. A density–exposure time curve does not possess a toe or threshold but passes straight through the origin where its slope is highest (Glocker and Traub 1921, Broili and Kiessig 1934). This is true for all emulsions except those of exceedingly fine grain (Lippmann emulsions) which latter have a toe for x rays and α-rays (Blau and Wambacher 1933); with these apparently the number of silver atoms liberated in each grain is not sufficient to make that grain developable. There is no evidence as to whether these emulsions are subject to reciprocity failure.

In a recent paper by Pelc (1945) the "grain quantum yield" of x rays has been investigated and it has been found that a single quantum of x rays absorbed is responsible for making several grains developable. So far we do not know whether it is secondary x rays or electrons which are responsible for the production of more than one developable grain per quantum.

This finding accounts for the peculiarly granular appearance of developed densities after exposure to x rays (Eggert and Schopper 1938) as compared with exposure to light, since the grains exposed by the secondary radiation must lie around the grain originally hit by the x-ray quantum (Herz, private communication).

Solarization occurs with x-ray exposures (Schaum 1904). Presumably x-ray exposures are comparable with exposures to light at very high intensities as evidenced by the Villard effect (§ 2.2.12), which like the Clayden effect, is probably due to the occurrence of internal latent image silver.

In a recent paper Hess (1943) has shown that the silver halide lattice becomes widened on exposure to x rays, contracting again as the exposure becomes sufficient to solarize the emulsion. The x-ray energy absorbed by the grains corresponds to the work done against the lattice forces. There is no suggestion here that it is this widening which constitutes the latent image, since in the case of light the energy absorbed would be insufficient to affect the crystal structure. These results have been contradicted by Brentano and Spencer (1946).

Attention has been given more recently to the problems connected with the direct recording of nuclear tracks in photographic emulsions. Most of the work has been devoted to the practical problems of recording in the special highly concentrated emulsions that have been evolved for this purpose (Powell, Occhialini, Livesey and Chilton 1946), or of producing these emulsions (Demers 1947), the sensitivity of which has been brought up sufficiently to be capable of recording tracks of quite weakly ionizing particles such as slow electrons (Berriman 1948). The specific latent-image problems involved have been discussed by Cüer (1947).

I wish to acknowledge with gratitude the continued help in this work received from Professor N. F. Mott, F.R.S., and especially his permission to quote from unpublished work. In compiling the review I had many helpful discussions with Mr. P. C. Burton, Dr. G. W. W. Stevens, and our colleagues in the Kodak

organization here and in Rochester. It is difficult, if not impossible, to acknowledge individually all the suggestions and ideas presented in this Report which is the result of discussions spread over several years.

REFERENCES

NOTE. No attempt has been made to quote in each instance the first worker in the field. Preference has often been given to references which were thought to be more readily accessible than the original papers.

ABEGG, R., 1897, *Ann. d. Phys., Lpz.*, **67,** 425 ; 1899, *Arch. wiss. Photogr.*, **1,** 15.

ABNEY, W. DE W., 1882, *Instruction in Photography* (London).

ALBERT, E., 1899, *Arch. wiss. Photogr.*, **1,** 285.

ARENS, H., 1931, *Z. wiss. Photogr.*, **30,** 127.

BAGDASSARIAN, C., 1944 a, *Acta Physicochim. U.R.S.S.*, **19,** 421 ; 1944 b, *Ibid.*, **19,** 436.

BARSHCHEVSKY, 1946, *Zh. Eksp. Teor. Fiz., U.S.S.R.*, **16,** 815.

BARTELT, O. and KLUG, H., 1934, *Z. Phys.*, **89,** 779.

BAUR, E., 1903, *Z. Phys. Chem.*, **45,** 613.

BERG, W. F., 1938 a, *Proc. Roy. Soc.* A, **164,** 79 ; 1938 b, *Trans. Faraday Soc.*, **34,** 889 ; 1939, *Ibid.*, **35,** 445 ; 1940, *Proc. Roy. Soc.* A, **174,** 559 ; 1942, *Ann. Rep. Progr. Chem.*, **39,** 49 ; 1943, *Trans. Faraday Soc.*, **39,** 115 ; 1945, *Phil. Mag.*, **36,** 337 ; 1946, *Photogr. J.*, **96** B, 154 ; 1947, *Ibid.*, **87** B, 112 ; 1948, *Trans. Faraday Soc.*, in press.

BERG, W. F., MARRIAGE, A. and STEVENS, G. W. W., 1941, *Photogr. J.*, **81,** 413.

BERG, W. F. and MENDELSSOHN, K., 1938, *Proc. Roy. Soc.* A, **168,** 168.

BERRIMAN, R. W., 1948, *Nature, Lond.*, in press.

BLAU, M. and WAMBACHER, H., 1933, *Z. wiss. Photogr.*, **31,** 243.

BLOCH, O. and RENWICK, F. F., 1920, *Photogr. J.*, **60,** 145.

BODENSTEIN, M., 1941, *Abh. Preuss. Akad. Wiss. Math.-nat. Kl.*, No. 19.

DE BOER, J. H., 1935, *Electron Emission and Absorption Phenomena* (Cambridge), p. 299.

BREIDO, I. I. and GOROKHOVSKI, Y. N., 1943, *J. Phys. Chem., U.S.S.R.*, **17,** 57 ; 1944, *Ibid.*, **18,** 152.

BRENTANO, J. C. M. and BAXTER, S., 1940, *Trans. Faraday Soc.*, **36,** 581.

BRENTANO, J. C. M. and SPENCER, L. VAN C., 1946, *Bull. Amer. Phys. Soc.*, **21,** 10.

BROILI, H. and KIESSIG, H., 1934, *Z. Phys.*, **87,** 425.

BROOKER, L. S., 1942, *Rev. Mod. Phys.*, **14,** 275.

BULLOCK, E. R., 1927, *Chemical Reactions of the Photographic Latent Image* (New York).

BURTON, P. C., 1946, *Photogr. J.*, **86** B, 62 ; 1948, *Ibid.* (in press).

BURTON, P. C., and BERG, W. F., 1946, *Photogr. J.*, **86** B, 1.

CABANNES, 1929, *La Diffusion moléculaire de la Lumière* (Paris), p. 83.

CAMERON, A. E. and TAYLOR, A. M., 1934, *J. Opt. Soc. Amer.*, **24,** 316.

CAREY LEA, M., 1918, *Colloides Silber und die Photohaloide* (Dresden : Steinkopf).

CARROLL, B. H. and HUBBARD, D., 1931 a, *Bur. Stand. J. Res., Wash.*, **7,** 219 ; 1931 b, *Ibid.*, **7,** 811.

CARROLL, B. H. and KRETCHMAN, C. M., 1933, *Bur. Stand. J. Res., Wash.*, **10,** 449.

CLARK, W., 1923, *Photogr. J.*, **63,** 230 ; 1924, *Ibid.*, **64,** 91 ; 1939, *Photography by Infra-red* (London), p. 83.

CÜER, P., 1947, *Sci. Industr. Photogr.*, **18,** 321, 353.

DEBOT, R., 1941 a, *Bull. Soc. Roy. Sci. Liège*, p. 90 ; 1941 b, *Ibid.*, p. 675 ; 1942 a, *Mem. Soc. Roy. Sci. Liège*, **6,** 91 ; 1942 b, *Bull. Soc. Roy. Sci. Liège*, p. 126 ; 1943, *Ibid.*, p. 647.

DEMERS, P., 1947, *Can. J. Res.*, **24** A, 223.

DERSCH, F. and DÜRR, H., 1937, *J. Soc. Inst. Pict. Engrs.*, **28,** 178.

EGGERT, J. and BILTZ, M., 1940, *Z. wiss. Photogr.*, **39,** 140.

EGGERT, J. and KLEINSCHROD, F. J., 1940, *Z. wiss. Photogr.*, **39,** 155.

EGGERT, J. and KÜSTER, A., 1936, *Kinotech.*, **23,** 381.

EGGERT, J. and NODDAK, W., 1921, *Sitzungsber. Preuss. Akad. Wiss.*, p. 631 ; 1923 a, *Ibid.*, p. 116 ; 1923 b, *Z. Phys.*, **20,** 299 ; 1924, *Ibid.*, **21,** 264 ; 1925, *Ibid.*, **31,** 922 ; 1927 a, *Ibid.*, **43,** 222 ; 1927 b, *Naturwissenschaften*, **15,** 57 ; 1929, *Z. Phys.*, **58,** 861.

EGGERT, J. and SCHOPPER, E., 1938, *Z. wiss. Photogr.*, p. 221.
EVANS, C. H. and HIRSCHLAFF, E., 1939, *J. Opt. Soc. Amer.*, **29**, 164.
EVANS, R. M. and HANSON, W. T., 1937, *Photogr. J.*, **77**, 497.
FAJANS, K., 1922, *Z. Elektrochem.*, **28**, 499.
FAJANS, K. and BECKERATH, 1921, *Z. phys. Chem.*, **97**, 478.
FAJANS, K. and FRANKENBURGER, W., 1923, *Z. phys. Chem.*, **105**, 255.
FAJANS, K. and STEINER, W., 1927, *Z. phys. Chem.*, **125**, 307.
FALLA, L., 1941, *Bull. Soc. Roy. Sci. Liège*, p. 395.
FARRER, W. J. G., 1936, *Photogr. J.*, **76**, 486.
FESEFELDT, H. and GYULAI, Z., 1929, *Nachr. Ges. Wiss. Göttingen. Math. Phys. Kl.*, II, p. 226.
FRANCK, J. and TELLER, E., 1938, *J. Chem. Phys.*, **6**, 861.
FRANKENBURGER, W., 1923, *Z. phys. Chem.*, **105**, 273.
FREUNDLICH, H., 1936, *Photogr. J.*, **76**, 49.
FRIEDRICH, W. and KOCH, P. P., 1914, *Ann. d. Phys., Lpz.*, **45**, 399.
GLOCKER, R. and TRAUB, W., 1921, *Phys. Z.*, **22**, 345.
GRIMLEY, T. B. and MOTT, N. F., 1947, *Trans. Faraday Soc.* (in press).
GURNEY, R. W. and MOTT, N. F., 1938, *Proc. Roy. Soc.* A, **164**, 151.
HALL, C. E. and SCHOEN, A. L., 1941, *J. Opt. Soc. Amer.*, **31**, 281.
HAMMER, R., 1922, *Phys. Rev.*, **20**, 189.
HARTUNG, E. J., 1924, *J. Chem. Soc.*, **125**, 2198.
HAUTOT, A. and FALLA, L., 1942, *Bull. Soc. Roy. Sci. Liège*, p. 298.
HAYNES, R. and SHOCKLEY, W., 1948, paper in *Report of a Conference on the Strength of Solids* (London : Physical Society). In press.
HERZFELD, K. F., 1923, *Z. phys. Chem.*, **105**, 329.
HESS, B., 1943, *Phys. Z.*, **44**, 245.
HILSCH, R. and POHL, R. W., 1930, *Z. Phys.*, **64**, 606.
HODGSON, M. B., 1917, *J. Franklin Inst.*, **184**, 705.
HOFFMANN, P. O., 1935, *Phys. Z.*, **36**, 650.
HUGGINS, M. L., 1943, *J. Chem. Phys.*, **11**, 412, 419.
HUSE, E. and MEULENDYKE, C. E., 1926, *Photogr. J.*, **66**, 306.
JAMES, T. H., 1939, *J. Phys. Chem.*, **43**, 701 ; 1943, *J. Chem. Phys.*, **11**, 338 ; 1944, *Ibid.*, **12**, 453 ; 1945 a, *J. Franklin Inst.*, **240**, 83 ; 1945 b, *Ibid.*, **240**, 229 ; 1945 c, *Ibid.*, **240**, 327 ; 1947, *Ibid.*, **243**, 235.
JAMES, T. H., BLAIR, J. M. and GERMANN, F. E. E., 1934, *J. Phys. Chem.*, **38**, 1023.
JAMES, T. H. and KORNFELD, G., 1942, *Chem. Rev.*, **30**, 1.
JAUSSERAN, C., 1943, *Sci. Industr. Photogr.* (2), **14**, 97.
JONES, L. A. and NELSON, C. N., 1940, *J. Opt. Soc. Amer.*, **30**, 93.
JOST, W., 1938, *Trans. Faraday Soc.*, **34**, 860.
KAMEYANA, N. and FUKAMOTO, 1939, *J. Soc. Chem. Industr. Japan*, **42**, 244 B (quoted by West and Carroll (1947 a)).
KAMEYANA, N. and MIZUTA, K., 1939, *J. Soc. Chem. Industr. Japan*, **42**, 426 B (quoted by West and Carroll (1947 a)).
KEMPF, A., 1937, *Z. wiss. Photogr.*, **36**, 235.
KIESER, H., 1928/9, *Z. wiss. Photogr.*, **26**, 1.
KOCH, E. and WAGNER, C., 1937, *Z. phys. Chem.* B, **38**, 295.
KOCH, P. P. and VOGLER, H. J., 1925, *Ann. d. Phys., Lpz.*, **77**, 495.
KÖGELMANN, F., 1894, *Isolierung der Substanz des latenten Bildes* (Graz).
KOLTHOFF, I. M. and O'BRIEN, A. S., 1939, *J. Amer. Chem. Soc.*, **61**, 3409.
KORNFELD, G., 1938, *J. Phys. Chem.*, **42**, 795 ; 1941, *J. Opt. Soc., Amer.*, **31**, 598.
VAN KREFELD, A., 1933, *Physica*, **1**, 60 ; 1934, *Z. wiss. Photogr.*, **23**, 222.
VAN KREFELD, A. and JURRIENS, H. J., 1924, *Physica*, **4**, 285.
KRON, E., 1913, *Publ. Astrophys. Obs, Potsdam*, **22**, 1.
LAINER, A., 1891, *Photogr. Korr.*, **28**, 12.
LANDAU, R., 1931, *Sci. Industr. Photogr.*, **2**, 1 ; 1934, *Ibid.*, **5**, 137.
DE LANGHE, 1936, *Photogr. Korr.*, p. 150.
LEERMAKERS, J. A., 1937, *J. Chem. Phys.*, **5**, 889.
LEERMAKERS, J. A., CARROLL, B. H. and STAUD, C. J., 1937 a, *J. Chem. Phys.*, **5**, 878 ; 1937 b, *Ibid.*, **5**, 893.

LEHFELDT, W., 1935, *Nachr. Ges. Wiss. Göttingen, Math. Phys. Kl.*, II, p. 170.
LESCYNSKI, W., 1926/7 a, *Z. wiss. Photogr.*, **24**, 261 ; 1926/7 b, *Ibid.*, **24**, 275.
LEVENSON, G. I. P., 1946, *Studies on the Action of the p-Methylaminophenol (metol)-cum hydroquinone developer* (London : Ph.D. Thesis) ; 1947, Paper read at XIth International Congress of Pure and Applied Chemistry (London).
LÖHLE, F., 1933, *Nachr. Ges. Wiss. Göttingen, Math. Phys. Kl.*, II, p. 271.
LÜPPO-CRAMER, H., 1921, *Kolloidchemie und Photographie* (Dresden) ; 1927, in Eder, *Handb. d. Photogr.*, **2**, Pt. 1, 526 (Halle) ; 1940, *Photogr. Industr.*, **38**, 271.
MARRIAGE, A., 1937, *Photogr. J.*, **77**. 619.
MAYER, J. E., 1933, *J. Chem. Phys.*, **1**, 327.
MEES, C. E. K., 1937, British Patents 498,031, 498,289, 498,291, U.S. Patents 2,075,046–8 ; 1942 a, *Theory of the Photographic Process*, p. 19 (New York) ; 1942 b, *Nature, Lond.*, **150**, 720.
MEIDINGER, W., 1932, in Hay, *Handb. d. Photogr.*, **5**, 271; 1935, *Phys. Z.*, **36**, 312 ; 1936, *Photogr. Industr.*, **34**, 1305 ; 1940, *Phys. Z.*, **41**, 277 ; 1943, *Ibid.*, **44**, 1.
MEIKLYAR, P., 1941, *C.R. Acad. Sci., U.R.S.S.*, **31**, 226.
MOORE, G. E., 1941, *Photogr. J.*, **81**, 27.
MOTT, N. F., 1946, *J. Phys. Radium*, **7**, 249.
MUTTER, E., 1928/9, *Z. wiss. Photogr.*, **26**, 193.
NAFE, J. E. and JAUNCEY, G. E. M., 1940, *Phys. Rev.*, **57**, 1048.
NUTTING, P. G., 1913, *Phil. Mag.*, **26**, 423.
OSTWALD, W., 1893, *Lehrb. d. allg. Chemie*, Vol. 2, Pt. 1 (Leipzig).
OTT, 1928, in Wien and Harms, *Handb. d. Exper. Phys.*, **7**, Pt. 2, 101.
PELC, S. R., 1945, *Proc. Phys. Soc.*, **57**, 523.
POHL, R. W., 1937, *Proc. Phys. Soc.*, **49** (extra number), 1.
POWELL, C. F., OCCHIALINI, G. P. S., LIVESEY, D. L. and CHILTON, L. V., 1946, *J. Sci. Instrum.*, **23**, 102.
RAWLING, S. O., 1926, *Photogr. J.*, **64**, 343.
REINDERS, W., 1911, *Z. phys. Chem.*, **77**, 213, 356, 677 ; 1934, *J. Phys. Chem.*, **38**, 784.
REINDERS, W. and HAMBURGER, L., 1932, *Z. wiss. Photogr.*, **31**, 32; 1933, *Ibid.*, **31**, 265.
RENWICK, F. F., 1920, *J. Soc. Chem. Industr.*, **39**, 156.
RITCHIE, M. and THOM, J. A., 1946, *Trans. Faraday Soc.*, **42**, 418.
ROSS, F. E., 1924, *The Physics of the Developed Photographic Image* (New York : D. Van Nostrand).
RUA, E., 1942, *C.R. Trav. Fac. Sci. Marseille* (1), **1**, 81 (cf. *Sci. Industr. Photogr.*, (2) **14**, 66 (1943)).
SABATTIER, 1927, in Eder, *Handb. d. Photogr.*, **2**, Pt. 1, 623.
SAVOSTJANOVA, M., 1935, *9th Congrès International de Photographie* (Paris), p. 94.
SCHAUM, K., 1904, *Arch. wiss. Photogr.*, **18**, 77.
SCHEFFER, F. E. C., 1907, *Brit. J. Photogr.*, **54**, 116, 271.
SCHEIBE, G., 1937, *Z. angew. Chem.*, **50**, 212.
SCHWARZ, G., 1933, *Photogr. Korr.*, **69**, 27.
SCHWARZ, G. and URBACH, F., 1932 a, *Photogr. Korr.*, **68**, 61 ; 1932 b, *Z. wiss. Photogr.*, **31**, 79.
SELWYN, E. W. H., 1939, *Phys. Soc. Rep. Progr. Phys.*, **5**, 191.
SHAPIRO, I. and KOLTHOFF, I. M., 1947, *J. Chem. Phys.*, **15**, 41.
SHEPPARD, S. E., 1925, *Photogr. J.*, **65**, 380 ; 1928, *Ibid.*, **68**, 397 ; 1929, *Ibid.*, **69**, 330 ; 1931, *Ibid.*, **71**, 331 ; 1942, *Rev. Mod. Phys.*, **14**, 303.
SHEPPARD, S. E. and HUDSON, J. H., 1927, *J. Amer. Chem. Soc.*, **49**, 1814 ; 1927/8, *Z. wiss. Photogr.*, **25**, 113.
SHEPPARD, S. E. and MEES, C. E. K., 1907, *Theory of the Photographic Process* (London).
SHEPPARD, S. E., TRIVELLI, A. P. H. and LOVELAND, R. P., 1925, *J. Franklin Inst.*, **200**, 51.
SHEPPARD, S. E., TRIVELLI, A. P. H. and WIGHTMAN, E. P., 1927, *Photogr. J.*, **67**, 281.
SHEPPARD, S. E., VANSELOW, W. and QUIRK, R. F., 1945, *J. Franklin Inst.*, **240**, 452.
SHEPPARD, S. E., WIGHTMAN, E. P. and QUIRK, R. F., 1934, *J. Phys. Chem.*, **38**, 817.
SILBERSTEIN, L., 1941, *J. Opt. Soc. Amer.*, **31**, 343 ; 1945, *Ibid.*, **35**, 93.
SILBERSTEIN, L. and TRIVELLI, A. P. H., 1938, *J. Opt. Soc. Amer.*, **28**, 441.
SILBERSTEIN, L. and WEBB, J. H., 1934, *Phil. Mag.*, **18**, 1.

STASIW, O. and TELTOW, J., 1941 a, *Ann. d. Phys., Lpz.*, **40**, 181 ; 1941 b, *Nachr. Akad. wiss. Göttingen, Math. Phys. Kl.*, pp. 93, 100, 110 ; 1944, *Ibid.*, p. 155.
STEINER, W., 1927, *Z. Phys. Chem.*, **125**, 275.
STERRY, J., 1904, *Photogr. J.*, **44**, 50.
STEVENS, G. W. W., 1939, *Photogr. J.*, **79**, 21 ; 1942, *Ibid.*, **82**, 42.
STEVENS, G. W. W. and NORRISH, R. G. W., 1938, *Photogr. J.*, **78**, 513.
STROCK, L. W., 1938, *Norske Vidensk. Skr. I. Mat.-Naturv. Kl.* No. 10.
SVEDBERG, T., 1920/1, *Z. wiss. Photogr.*, **20**, 36 ; 1922, *Photogr. J.*, **62**, 186, 310.
SVEDBERG, T. and ANDERSON, H., 1921, *Photogr. J.*, **61**, 325.
TOLLERT, H., 1929, *Z. phys. Chem.* A, **140**, 355.
TOY, F. C., 1921, *Photogr. J.*, **61**, 417 ; 1928, *Proc. 7th International Congress of Photography* (London), p. 14.
TOY, F. C. and HARRISON, G. B., 1930, *Proc. Roy. Soc.* A, **127**, 613.
TRILLAT, J. J. and MERIGOUX, R., 1936, *J. de Phys.*, **9**, 497.
TRIVELLI, A. P. H., 1929, *J. Franklin Inst.*, **207**, 765 ; 1946 a, *Ibid.*, **241**, 1 ; 1946 b, *Ibid.*, **241**, 85.
TUBANDT, C., 1932, in Wien and Harms, *Handb. d. Exper. Phys.*, **12**, Pt. 1.
UNGAR, G., 1937, *Z. Phys.*, **106**, 322.
VASSY, A., 1947, *Sci. Industr. Photogr.* (2), **18**, 65.
VASSY, A., PINOIR, R. and VASSY, E., 1945, *Sci. Industr. Photogr.*, **16**, 249.
VASSY, A., SCHONNE, P., and VASSY, E., 1947, *C.R. Acad. Sci., Paris*, **224**, 910.
VASSY, A. and VASSY, E., 1947, *C.R. Acad. Sci., Paris*, **225**, 55.
DE VAUCOULEUR, G. and VIOLLET, R., 1947, *Sci. Industr. Photogr.* (2), **18**, 97.
VOLMER, M., 1932, *Trans. Faraday Soc.*, **28**, 359.
WATERHOUSE, J., 1890, *Brit. J. Photogr.*, **37**, 601, 613.
WEBB, J. H., 1933 a, *J. Opt. Soc. Amer.*, **23**, 157 ; 1933 b, *Ibid.*, **23**, 316 ; 1935, *Ibid.*, **25**, 4 ; 1936, *Ibid.*, **26**, 12 ; 1941, *Ibid.*, **31**, 559 ; 1942, *Ibid.*, **32**, 299.
WEBB, J. H. and EVANS, C. H., 1938, *J. Opt. Soc. Amer.*, **28**, 249 ; 1940 a, *Ibid.*, **30**, 445 ; 1940 b, *Photogr. J.*, **80**, 188.
WEINLAND, C. E., 1927, *J. Opt. Soc. Amer.*, **15**, 337 ; 1928, *Ibid.*, **16**, 295.
WEST, W. and CARROLL, B. H., 1947 a, *J. Chem. Phys.*, **15**, 529 ; 1947 b, *Ibid.*, **15**, 539.
WIGHTMAN, E. P. and QUIRK, R., 1927, *J. Franklin Inst.*, **203**, 261.
WILSEY, R. B., 1921, *Phil. Mag.*, **42**, 262; 1925, *J. Franklin Inst.*, **200**, 739.
WILSON, 1907, *Ann. d. Phys., Lpz.*, **22**, 107.
WOLFE, R. N. and BURROWS, R. S., 1947, *P. S. A. J.*, **13**, 554.
WOOD, R. W., 1903, *Phil. Mag.*, **6**, 577.
ZENKER, 1868, *Lehrb. d. Photochromie* (quoted by Sterry (1904)).
ZIMENS, K. E., 1946, *Arkiv f. Kemi, mineralogi och Geologi*, **23** A, 1.

The following books and reviews will be found useful to those interested in further details:—

DE BOER, 1935, *Electron Emission and Absorption Phenomena* (Cambridge).
BROOKER, 1942, *Rev. Mod. Phys.*, **14**, 275 (absorption and resonance in dyes).
EUCKEN, 1944, *Lehrbuch d. chem. Phys.*, **2**, 2 (Leipzig) (solid state).
GURNEY, 1936, *Ions in Solutions* (Cambridge).
HAY, 1932, *Handb. d. Photographie*, **5** and **6** (Vienna) (very full and careful references).
HERZFELD and SKLAR, 1942, *Rev. Mod. Phys.*, **14**, 224. (polymethine dyes).
JAMES and KORNFELD, 1942, *Chem. Rev.*, **30**, 1 (development).
MEES, 1942, *Theory of the Photographic Process* (New York).
MOTT and GURNEY, 1940, *Electronic Processes in Ionic Crystals* (Oxford).
SHEPPARD, 1942, *Rev. Mod. Phys.*, **14**, 303 (environment and spectra of dyes).
STENGER and STAUDE, 1944, *Fortschritte d. Photographie* (Leipzig) (article on the latent image by MEIDINGER).

[*Note added in proof.* Quite recently, the mechanism of photolysis of silver halides suggested by Gurney and Mott (1938) (see § 5.1), and in particular the part played by electronic movements, has found a striking confirmation (Haynes and Shockley, 1948). On applying alternate flashes of light and pulses of D.C. potential to a silver chloride crystal, the discoloration due to the print-out silver is moved from the illuminated area in the direction towards the anode.]

THE MECHANISM OF THE THERMIONIC EMISSION FROM OXIDE COATED CATHODES

By

H. FRIEDENSTEIN *, S. L. MARTIN AND G. L. MUNDAY †,
Material Research Laboratory, Philips Electrical Ltd., Mitcham, Surrey

CONTENTS

	Page
§ 1. Introduction	298
§ 2. The main characteristics of oxide cathodes	301
2.1. Activation and de-activation	301
2.11. Activation	301
2.12. De-activation	304
2.2. Time changes exhibited by the fully activated cathode	305
2.21. Time changes on drawing of current	305
2.22. Life failure	307
2.23. Temperature effects	308
2.3. Lack of saturation	308
§ 3. Processes determining the emission	309
3.1. Introduction	309
3.2. Conduction	310
3.21. Experimental data	310
3.22. Theory of the conduction process	312
3.3. Contact phenomena and the core coating interface	317
3.31. Contact phenomena	317
3.32. Rôle of the core metal and the interface	318
3.4. Diffusion	319
3.5. Evaporation	321
§ 4. General theories of emission	322
4.1. Introduction	322
4.2. The Wilson model	322
4.3. de Boer's model	326
4.4. The model of Uehara and Takahari	327
4.5. General considerations	328
§ 5. Other phenomena occurring with oxide coated cathodes	330
5.1. Luminescent properties	331
5.2. Thermionic emission under pulsed conditions	331
5.3. Secondary electron emission under electron bombardment	335
§ 6. Conclusions	336
Appendix : List of recurrent symbols used	338
References	339

§ 1. INTRODUCTION

THE emission of charged particles, and in particular of electrons, from hot bodies is a well-known phenomenon. It is due to thermal excitation of the electrons within the body to a state of sufficiently high energy to overcome the energy-barrier set up by image forces and polarization of the atoms at the surface. If the hot body is surrounded by a vacuum and no external field is applied, a negative space-charge is set up round the emitter and most

* Now at British Oxygen Company, Research and Development Department, Morden Factory Estate, London.

† Now at Physics Department, University of Birmingham.

of the electrons will fall back into the emitter. Owing to the Maxwellian velocity distribution of the emitted electrons, a few can overcome the space-charge, producing, with no external positive field (or even with a small negative one), a small 'starting current' otherwise referred to as the 'retarding field emission'. In order to collect more of the electrons, the space-charge must be overcome by applying a suitable positive electric field. So long as this field is small, the current will be voltage-dependent, and we have what are known as 'space-charge-limited' or 'unsaturated' conditions. 'Saturation' is obtained when a sufficiently large potential is applied to overcome the space-charge completely. All the emitted electrons are then collected on the anode and, theoretically at least, the current is independent of the field. In practice, complete saturation is never obtained. In the case of metals, the so-called 'saturation current' depends on the field according to a half-power law and this has been satisfactorily accounted for by theory (Schottky, 1914). Thus, for metals at least, the saturation current for zero field can always be calculated from the observed current, and this 'zero-field emission' is the characteristic property of the emitter, which is generally termed 'the emission'. Unless otherwise specifically stated, this review is concerned with the saturated emission.

It is obvious that this emission current must depend on the temperature of the emitter. It is often termed 'temperature-limited current', as opposed to the space-charge-limited current, which depends on the field. The temperature dependence of the emission is represented by Richardson's equation (Richardson, 1914; Richardson and v. Laue, 1918)

$$i_s^0 = AT^2 \exp(-e\phi/kT), \qquad \ldots\ldots(1)$$

where i_s^0 = saturated emission, in amp/cm^2 at zero field; T = temperature, in °K; A = constant (known as the emission constant); ϕ = constant = work function, in ev.; k = Boltzmann's constant; and e = electronic charge.

This general relation applies without reservations to metals. The work function ϕ is a characteristic constant for any pure metal, depending only on the surface conditions in the case of gaseous or other contaminations. Following theoretical derivations of the emission equation, A should be a universal constant, calculated to be 120 amp. cm^{-2} deg^{-2}, and it has been found to be of the correct order of magnitude for most metals.*

By plotting $(\log i_s^0)/T^2$ against $1/T$ a value is obtained for the experimental work function ϕ_e (which, though approximating to, may still be significantly different from the work function ϕ_0 at 0° K.†), together with a value for A.

If ϕ is calculated by (1) from i_s^0 at a particular temperature with the theoretical

* A complete theoretical treatment leads to a value for $A = \bar{D}A_0 = (1-\bar{r})A_0$ where A_0 is the theoretical constant of 120 amp. cm^{-2} deg^{-2}; $\bar{r}$ is the mean reflection coefficient of electrons reaching the surface with sufficient energy to overcome the surface potential barrier, i.e. the fraction of electrons internally reflected; $\bar{D}$ is thus the mean transmission coefficient. For most pure metals, however, it appears that $\bar{r}$ is, to a first approximation, negligible. For a detailed treatment of the reflection and transmission coefficients see Reimann (1934), especially Chapter V.

† Theoretical derivations which neglect the specific heat of the internal electrons at all temperatures, i.e. which assume a temperature coefficient $d\phi/dT = 0$, give emission equations similar to equation (1) but with the constants A_0 and ϕ_0. If in the corrected equation $d\phi/dT$ is within experimental error independent of T, then it can be shown that $d\phi/dT$ will appear in the A-term, and only ϕ_0 in the exponential temperature-dependent term (see Reimann 1934, p. 267). In that case $\phi_e = \phi_0$; but as $d\phi/dT$ may be significantly temperature dependent, the distinction between ϕ_e and ϕ_0 will be maintained here.

A value, the value ϕ_T so obtained applies to the temperature of measurement; the temperature coefficient of ϕ for most metals is, however, quite small.

Matters are not so simple when we come to consider emitters other than pure metals. One of the most important emitters for practical purposes is the oxide cathode, which consists of a metallic core—either a filament or a metal-sleeve which can be heated indirectly—with an oxide coating, usually consisting of one or more of the oxides of the alkaline-earth metals. This type of composite emitter, which was first discovered by Wehnelt in 1903, begins to emit electrons at considerably lower temperatures than pure metals or even the more active 'contaminated metals' of the tungsten-caesium type, and thus constitutes a more active and economical source of electrons.

As in the case of metals, the emission is an exponential function of the temperature. But there the resemblance ends. In the first place, it is difficult to define 'the emission' of an oxide cathode. The emission current does not saturate to the same extent as that from pure metals, and Schottky's relation is not obeyed. Furthermore, an oxide cathode exhibits changes in activity (determined as the emission current at a particular temperature and anode voltage) with time, following upon and accompanying changes in temperature, or anode potential, or in the current drawn from it. Thus, 'Richardson lines' (as the plots of $(\log i_s)/T^2$ against $1/T$ are termed) of widely differing slopes and intercepts can be obtained for the same cathode under slightly varying measuring conditions or methods of pretreatment, and it becomes difficult to define the work function or A value of any particular oxide. Oxide cathodes require 'activation' in order to reach maximum emission, and the ultimate activity depends on the history of the cathode. Finally, the work function appears to have a larger and more variable temperature coefficient. Thus it is not surprising that the results of experimental investigations are widely divergent and that there is considerable disagreement regarding the general properties of oxide cathodes and the mechanism of electron emission from them.

It is not proposed to give a detailed survey of all the information available in the literature regarding all the properties of oxide cathodes and the technique of their production and measurement. Most of the data have been collected by Blewett (1939a), and a general account of the characteristics of these emitters is available in the older reviews and textbooks (Dushman, 1930; Reimann, 1934; de Boer 1935a and 1937), and in the recent books on oxide cathodes by Hermann and Wagener (1943, 1944).* The main processes involved in the preparation of oxide cathodes are briefly :—

(*a*) Preparation of powders (carbonates) and suspensions thereof (generally containing an organic binder) suitable for coating.

(*b*) Coating—by cataphoresis, spraying, dipping or dragging—of filaments or sleeves.

(*c*) Removal of binder and decomposition of carbonate to oxide by suitable heat treatment while exhausting.

(*d*) Subsequent activation after completion of the exhaust process, by temperature treatment and/or current 'forming'.

* This Report was written before these volumes became available, but has been revised to include the more important aspects of the recent German work reported therein.

This Report is written to serve a two-fold purpose. On the one hand, it is intended to collect and correlate, as far as possible, the experimental results, which are at present distributed over a large number of publications, regarding those properties of the oxide cathode which appear to be most relevant to the elucidation of the emission mechanism; this aspect is covered in § 2 and para. 3.21. On the other hand, the main theories of the emission mechanism are compared and critically examined and an attempt is made to establish a working picture of the processes involved. In § 3, the basic processes which determine the emission are discussed. The main prevailing theories of the emission are critically reviewed and correlated in § 4. In § 5 recent information on other phenomena occurring with oxide cathodes which may have some bearing on the normal thermionic emission is briefly reviewed.

The symbols used by various authors have been standardized to conform as far as possible with those recommended in the Report of the Joint Committee of the Chemical Society, the Faraday Society and the Physical Society (1937); a list is given in the appendix on page 338.

§ 2. THE MAIN CHARACTERISTICS OF OXIDE CATHODES

2.1. *Activation and de-activation*

2.11. *Activation*

Oxide-coated cathodes will reach full activity only after a process of 'activation' or 'forming'. This may consist solely of thermal treatment (thermal activation) or the cathode may be held at a moderate temperature, a high potential applied simultaneously and current drawn (current activation or forming). It is generally assumed that in the process of activation free alkaline earth metal is formed. At the time of publication of Dushman's review (1930), there were two opposing schools of thought regarding the mechanism of liberation of the free metal, which postulated that :—

1. The barium (and/or strontium and/or calcium) is produced by thermal reaction between the oxide and the core-material.
2. The barium is produced by electrolysis of the oxide, during the current activation stage.

The first view (Koller, 1925; Lowry, 1930; Gehrts, 1930), was based on evidence of cases where purely thermal treatment is sufficient for complete activation. Gehrts (1930) considered that activation takes place in two stages. In the first stage, barium is produced by thermal dissociation of the barium oxide—(a theory which has been completely disproved by Villars and Dushman (1930) by estimation of the dissociation vapour pressure of BaO)—and also by reaction with reducing substances in the core, such as Ir, Rh, Ni. During the second stage, the liberated oxygen is completely removed from the cathode. The second view was put forward by earlier workers (Becker, 1929 and 1931; Detels, 1927; Espe, 1927 a and b; Rothe, 1926) on the evidence that oxygen is liberated during activation and that activated cathodes are 'poisoned' by oxygen. Reimann and Murgoci (1930) supported this view on the basis of the observed temperature-dependence of the conductivity (cf. § 3.21). Dushman (1930) suggested that the exact mechanism of activation may actually vary with the core-material and the manner in which the oxide is deposited on the core.

Observations of later investigators confirm the suspicion that the activation mechanism depends on the core-material. Complete activation of oxide-coated nickel cathodes has been obtained by heat-treatment only (Maddock, 1935; Heinze and Wagener, 1936). Benjamin's observation (1935) that aluminium, titanium and magnesium in the core have a beneficial effect on the emission—which incidentally confirms Lowry's observations with Konel filaments (Lowry, 1930)—is explained on the hypothesis that barium is formed by a thermal reaction between the oxide and the reducing components of the core. On the other hand, Berdennikowa (1932) showed that electrolytic decomposition of the oxide takes place during activation of a cathode consisting of a Pt–Ir filament coated with Ba, Sr and Ca oxides. Experiments by G. L. Tawney at G.E.C. (Blewett, 1939a) indicate that cathodes with cores of spectroscopically pure Pt cannot be activated thermally, which substantiates the comparative results for nickel and platinum cores reported by Heinze and Wagener (1936).

The possibility of reduction by carbon traces formed during decomposition from the binder or by carbon monoxide from the binder or carbonates has frequently been raised (Blewett, 1939a). Huber (1941) found that the final state of activation thermally attainable with BaO–Ni cathodes was independent of the presence of a binder. On the other hand, oxide cathodes prepared from barium peroxide coatings gave lower emission values than those from carbonates when thermally activated at equivalent temperatures, and this can be considered as evidence for reduction by carbon monoxide from the carbonate. It must, however, be pointed out that this last result may be misleading because of possible complications with the peroxide coatings (for instance variations in the true temperatures with the differently prepared oxides). Moreover, in view of the relatively inefficient thermal activation attainable with carbonates on platinum cores, reduction by carbon monoxide can only be of secondary importance.

The optimum temperature for thermal activation depends not only on the method of preparing the oxide coating, but also on its composition (Huber, 1941); it increases for instance in the order BaO, SrO, CaO (as does the heat of formation of the oxides) for the pure oxides, and varies regularly, but non-linearly, between the extremes for a two component mixture as shown in figure 1.

Eisenstein (1946a) has recently shown by x-ray diffraction studies on equimolecular barium strontium oxide-coated nickel cathodes that the maximum temperature attained during thermal activation not only affects the rate of formation of true solid solutions and also the rate of growth of individual crystallites, but also limits their size, which reaches a maximum in a comparatively short period of time. It is not yet clear, however, how these considerations affect the state of activation of the cathode.

The amount of electrolysis taking place during the activation has been calculated by Isensee (1937), from the amount of oxygen liberated (cf. § 3.21). Berdennikowa (1932) and Fritz (1935) determined the quantity of free barium in the coating by chemical methods. According to Fritz, most of the excess barium is formed during the preparation of the cathode and only a small percentage of the total amount is liberated during the 'activation'. Both Berdennikowa's and Fritz's results were criticized by Prescott and Morrison (1938), on the grounds that, while Berdennikowa's method fails to distinguish

between free metal in the cathode and elsewhere in the tube, Fritz's microchemical method, although selective, is not sensitive enough.

Current activation is undoubtedly responsible for the formation of some free barium and almost entirely so with core metals possessing low reducing properties (platinum). However, recent results (Hermann and Wagener, 1944) indicate that its more important rôle with other core materials (with which free barium will already have been formed by the necessary previous temperature

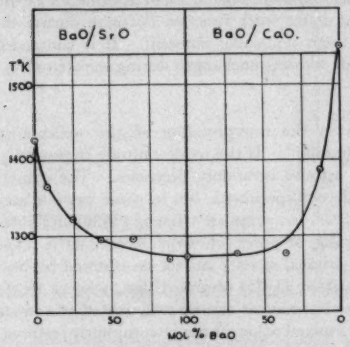


Figure 1. Optimum activation temperatures for (sprayed) oxide mixtures on nickel cathodes (Huber, 1941). True temperatures calculated from brightness temperature with $\epsilon_\lambda(\lambda=0{\cdot}665\ \mu)=0{\cdot}25$ for all compositions. Each point is the mean of a number of tubes.

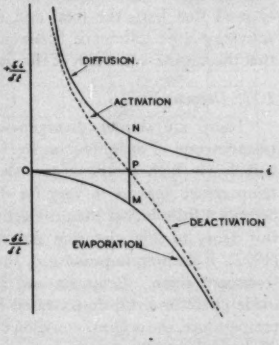


Figure 2. Assumed curves for rate of change of emission for one particular temperature (Maddock, 1935, figure 2).

treatment) is to 'reactivate' the cathode which becomes poisoned by traces of gas released from the anodes as soon as a current of any magnitude is drawn.

It is clear from the observations outlined above that the activation is a 'structure-sensitive' process, that is to say, it depends on the composition and previous history of the cathode. It is therefore not surprising that there are few and conflicting data regarding the variation of the emission-constants during the activation.

Espe (1927) observed that if the anode-voltage is raised after activating at any particular value, the constant A in the general equation (1) increases, whereas ϕ_0 remains constant. Conversely, the observed decrease in emission due to 'flashing' at a high temperature is attributed to a decrease in the A value. Detels (1927) stated that during the activation both A and the work function ϕ_0 decrease, but A does so more rapidly. Kniepkamp and Nebel (1932) also observed an initial decrease of both A and ϕ_0 during re-activation of a cathode after a decay in emission caused by drawing current (cf. paragraph 2.21); but the work function eventually became constant and A increased. Heinze and Wagener (1936) attribute the activation to a decrease in the 'inner work function', that is to say, the energy necessary to excite the electrons to the conduction-band of the oxide (cf. §3.22). They observed practically no change in the electron-optical image

of the cathode during the course of the activation, and concluded that the increasing emission cannot be due to any increase in the emitting area, except possibly to a small extent in the initial stages of the activation.

The results of later work (Heinze and Wagener, 1936) involving a contact potential method for measuring the work function, support the assumption that the increase in emission is due entirely to a decrease in the work function, which may amount to ~0·5 ev. By measuring the temperature-coefficient of the conductivity simultaneously with the emission, Nisibori and Kawamura (1940) showed that both the total and the inner work function decrease during the activation, the difference between them remaining constant. It is concluded that the surface-condition of the oxide remains unchanged during the activation.

2.12. *De-activation*

There are similar divergencies in the interpretation of the well-known phenomenon of de-activation by 'flashing'. If the oxide cathode is glowed at sufficiently high temperatures, the activity invariably decreases. The critical temperature appears to vary for different specimens, but in some cases it may be only a little higher than the activation temperature. Lowry (1930) attributes this decay to slow sintering and closing of pores (cf. § 4). According to Espe (1927), if flashing is prolonged, the original activity cannot be restored by heat treatment alone. Benjamin and Rooksby (1933) observed that, while a single oxide (BaO or SrO) de-activated in this manner can be re-activated at a lower temperature, the original emission of a mixed oxide cannot be completely restored. They attribute the loss of activity to evaporation of barium oxide (or possibly barium only).

Maddock (1935) observed that the rate of de-activation increases with increasing temperature and activity. The higher the de-activation temperature, the lower is the final emission. The rate of re-activation also increases with increasing temperature, but decreases as the activity increases. The final activity is higher, the lower the re-activation temperature. Maddock attributes this phenomenon to the establishment of an equilibrium between evaporation of barium from the surface and diffusion of barium towards the surface. The full curves in figure 2 show the assumed variation of the rates of evaporation and diffusion with increasing thermionic activity. The dotted curve, which represents the sum of the two, corresponds well to the experimentally observed rate of change in emission ($\delta i/\delta t$) as a function of the current i. At the point P, where the rate of diffusion is equal to the rate of evaporation, $\delta i/\delta t$ is zero, i.e. an equilibrium state is established. It is pointed out that the slope at this point is not zero, as might have been expected, a fact which is borne out experimentally.

According to Fan (1943), both work function and A value increase after high temperature flashing. He attributes the fact that he could not re-activate cathodes completely after repeated flashing at very high temperatures to poisoning by oxygen. Figure 3 shows the relationship between the work function ϕ_0 and $\log A$.

The effects of oxygen poisoning have been observed to be similar to those of high temperature flashing. Benjamin and Rooksby (1933) observed no recovery for poisoned mixed oxides, but the activity of single oxides could be completely restored by heating them to temperatures near the melting-point

of nickel and then re-activating at a lower temperature. Heinze and Wagener (1936, 1938) were able to re-activate poisoned barium oxide–Ni cathodes thermally at ~1250° K. They observed no effect of oxygen poisoning on the electron optical image; there was no noticeable change in the distribution or size of the emission centres at oxygen pressures ~10^{-5} mm. Hg, even though this pressure was sufficient to lower the saturated emission; after attack at ~10^{-3} mm. Hg the cathode temperature had to be raised by ~100° C. to reproduce the original image. This they interpreted as evidence for an 'outer' and 'inner' poisoning.

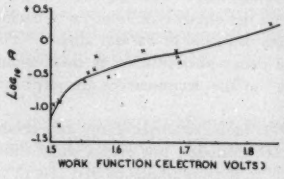


Figure 3. Relation between work function and log A (Fan, 1943, figure 9).

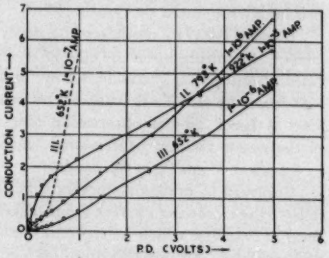


Figure 4. Variation of conduction current with P.D. (Reimann and Treloar, 1931, figure 3).

Another phenomenon which should be mentioned in this connection is the effect of 'idleness'. Here, the experimental evidence is conflicting. Espe (1927) states that, when a sealed valve is left 'idle' at room temperature, the activity of the cathode (consisting of oxide-coated Pt–Ir) decreases. He observed a decrease in the A value, which he attributed to recombination of barium with oxygen or back-diffusion of barium from the surface. According to Espe, the cathode can be re-activated by heating. On the other hand, Maddock (1935) observed an increase in activity by up to 10% following upon 'idleness', for barium strontium oxide-coated nickel cathodes. He offers no explanation for this effect.

2.2. *Time changes exhibited by the fully activated cathode*

Introduction

When a current is drawn from a fully activated cathode, a rapid decay from the initial value is observed, which gradually slows down until a final steady value is reached. The effect is probably due either to the removal of surface barium by electrolysis or to the deposition of oxygen on the surface, with the resultant oxidation of the active barium. From the experimental evidence (see § 2.21) it is difficult to decide which of the two mechanisms is the most important. The 'short time decay' accompanying the drawing of current under pulsed conditions (Sproull, 1945) is treated separately under pulse effects in § 5.2.

2.21. *Time changes on drawing of current*

With thermally activated cathodes Becker (1929) observed an increase in both ϕ_0 and A during the decay, which he attributes to the formation of a negative

layer on the surface by oxygen. Kniepkamp and Nebel (1932), who recorded the decay by an oscillographic method, find that during the initial stages ϕ_e remains constant while A decreases; then ϕ_e and A increase simultaneously. They attribute the latter part of the effect to the deposition of oxygen on the surface, but offer no explanation for the initial effect. In agreement with Becker, they find that the original activity can be recovered by glowing at a lower temperature. de Boer (1935 a, p. 364) attributes the initial decrease in A observed by Kniepkamp and Nebel to a temporary impoverishment of the surface in barium, as previously suggested by Gehrts (1930).

Blewett (1939 b) considers the decay to be a volume rather than a surface effect. Following upon the change in conditions due to drawing current, a new equilibrium state is gradually set up, in which the electrolytic removal of barium ions from the surface is balanced by outward diffusion of barium atoms. This view is based on the observation that the decay-phenomenon is independent of the anode voltage. Occasional 'humps' in the decay-curves are attributed to the effect of gas from the anode.

Contrary to previous workers, Fan (1943) failed to observe any appreciable decay when drawing current from nickel-core cathodes, but only a slight initial drop in emission which he attributes entirely to the simultaneous decrease in the temperature. During this drop, both A and ϕ_e decrease slightly. The temperature drop is due to the cooling effect of the emitted electrons, this being followed rapidly by a heating effect as current is drawn. His observations are completely substantiated by Hermann and Wagener (1944, p. 218), who find little or no change with well activated cathodes (Ni), either in the saturation (830°–935 K.) or space-charge-limited (660°–710° K.) regions provided rigid degassing schedules are adhered to.

In the light of the recent work under more carefully controlled experimental conditions, it is thus most probable that that type of sharp decrease in activity which is accompanied by a subsequent increase almost to the original value is ascribable to poisoning by gases released from anode or grids by the impinging electrons, followed by re-activation as current continues to be drawn.* More permanent effects are probably due to temperature changes accompanying the drawing of electrons from the cathode (see Reimann, 1934; Hermann and Wagener, 1944). The cooling effect Δq_e (watts cm^{-2}) due to a sudden removal of electrons can be represented by the equation:

$$\Delta q_e = i_s\left(\phi + \frac{2kT}{e}\right) \qquad \ldots\ldots(2a)$$

in the saturated current region, ϕ being the work function at the temperature T.

* Further convincing evidence for this has recently been given by Jacobs (1946) and by Hamaker Bruining and Aten (1947). Jacobs found that when an oxidized metal surface was bombarded with electrons from an oxide cathode, the cathode emission was poisoned at a certain critical bombarding potential which was approximately equal to the heat of formation of the particular metal oxide. Hamaker, Bruining and Aten described a similar poisoning effect when films of metal halides near the cathode were bombarded, and showed that such deleterious films of barium chloride can be produced because of evolutions of HCl from the glass under certain conditions. These observations further emphasize the difficulty of eliminating subsidiary effects due to uncontrollable contamination of the cathode when studying the activation and de-activation processes.

Because of the relatively high resistance of the oxide coating there will also be a heating effect Δq_h (watts cm^{-2}) given by

$$\Delta q_h = (i_s)^2 R_c, \qquad \ldots\ldots(2b)$$

where R_c is the ohmic resistance per cm^2 cathode surface. The total heat change affecting the temperature is thus

$$\Delta q_h - \Delta q_e = i_s\left(i_s R_c - \phi - \frac{2kT}{e}\right), \qquad \ldots\ldots(3)$$

the sign of which depends on the relative values of T, and of R_c and ϕ, both of which may vary for differently prepared cathodes. The observations previously cited can all be explained by suitable assumption of values of $\Delta q_h - \Delta q_e$ greater than, equal to, or less than zero, together with the fact that the rate of attainment of equilibrium accompanying the change Δq_h can be expected to be slower than in the case of the change Δq_e.

2.22. *Life failure*

Little quantitative information is available regarding the ultimate decrease in activity of oxide cathodes during life. Lowry (1930) attributes it to the same causes as the decay due to high temperature flashing, namely, sintering of the oxide layer. A similar interpretation is given by Reimann and Murgoci (1930). On the basis of Benjamin and Rooksby's experiments on high temperature flashing, Reimann (1934) considers that, in the case of mixed oxides, the 'life-failure' is due to preferential evaporation of barium oxide. It is well established, that the rates of evaporation of the alkaline earth oxides increase from CaO to BaO (Claasens and Veenemans, 1933; Burgers, 1933). Electron diffraction experiments (Gärtner, 1935; Darbyshire, 1938; Huber and Wagener, 1942) lead to the conclusion that the top layer of the activated mixed oxide cathode consists only of the least volatile of the oxides. The experiments carried out by Hull (1939) appear to support the view that evaporation of the oxide itself is largely responsible for the life-failure. He found no decrease in the emission properties after three years' life test in the case of a specially designed valve, in which the barium oxide evaporating from the cathode is replenished from a 'dispenser cathode' consisting of a $BaO.Al_2O_3$ eutectic on molybdenum.

According to Clausing (de Boer, 1935a, p. 358) the A value decreases during this decay, while ϕ_e remains unchanged. Both de Boer (1935a) and Blewett (1939a) take this to indicate that the decay is due to a failure in the supply of free alkaline earth metal. Veenemans (1943) observed a gradual increase of both the work function and the A value. He puts forward an explanation on the basis of de Boer's theory of the emission (cf. § 4) for the observed inter-dependence of ϕ_e and A, at different stages of activity. It is assumed that the ionization energy of the alkaline earth metal atoms adsorbed on the surface, as distinct from free atoms, varies from atom to atom, and that a small number of atoms will have a relatively small ionization energy. These constitute the most 'active' atoms which, according to de Boer (1935a), alone determine the work function. Veenemans shows that the 'effective' work function and A value depend on the proportion of active atoms on the surface. If this proportion changes, there will be a linear relationship between $\log A_{\text{eff.}}$ and $\phi_{\text{eff.}}$, as observed

experimentally. On this basis, the increase in log A and ϕ_0 during life is attributed to the recombination of barium at the more active points of the surface with oxygen, and to the consequent increasing contribution of the less active emission-centres to the effective work function.

2.23. *Temperature effects*

Another characteristic phenomenon which should be considered under the heading of 'time-changes' is the slow change in the emission following upon a change in temperature. When the temperature of the cathode is dropped suddenly, the emission drops to an initial lower value and then rises rapidly at first and then more slowly, until finally a steady value is reached. This was first reported by Davisson and Germer (1924). Reimann (1934, p. 189) attributes this effect to a temperature dependence of the work function. Blewett (1939 a, p. 843) points out that an effect of this type is to be expected if the steady emission is recognized as characteristic for a state of equilibrium between the diffusion of barium atoms and the electrolysis. If the temperature is changed suddenly, this equilibrium is disturbed and a variation in the current results, until a new equilibrium, corresponding to the new temperature, is established.

2.3. *Lack of saturation*

All emitters exhibit some increase in current with increasing anode potential in the region of so-called 'saturation'. This is primarily caused by a lowering of the work function due to the applied field. The theory of this effect has been worked out by Schottky (1914). A relationship is obtained of the form

$$i_s = i_s^0 \exp\left[e^{3/2}F^{1/2}/kT\right], \qquad \ldots\ldots(4)$$

where i_s^0, i_s, = saturated currents at zero field and anode voltage V respectively.

F = field strength at surface,

= V_A/d for flat electrodes where V_A is the anode voltage and d the distance anode–cathode surface.

This agrees fairly satisfactorily with experimental results for pure metals. Certain composite emitters, however, show a very much greater increase in current in moderate fields than can be accounted for by Schottky's equation. Langmuir (1920) first suggested that, for thoriated tungsten and similar systems, this can be accounted for by a non-uniform variation of the work function over the surface. He put forward a 'theory of patches', which was later developed by several investigators (Richardson and Young, 1925; Becker and Mueller, 1928; Reynolds, 1930; Compton and Langmuir, 1930; Linford, 1930 and 1931; Becker, 1935; Rose, 1936). Linford (1931) showed that photoelectric data can also be accounted for by assuming a 'checker board' distribution as suggested by Compton and Langmuir (1930), with square patches of side $\sim 10^{-4}$ cm., and that the contact potential difference between patches depends on the amount of adsorbed material. Becker (1935) arrived at similar conclusions on the basis of a more general theory.

It has been suggested that the patch theory may be applicable to oxide cathodes as well as other complex emitters. Huxford (1931) showed that the photoelectric threshold of oxide cathodes moves to longer wavelengths as the

applied field increases. Linford (1933, p. 51) showed that the observed fields in this case could be accounted for by small scale fluctuations in the work function (patches $\sim 10^{-5}$ cm.) superimposed on larger patches ($\sim 10^{-4}$ cm.). Reimann (1934, p. 207) suggests that a 'sufficiently large scale lack of uniformity of the adsorbed barium film' on the oxide would manifest itself as patches of different work function and thus be responsible for the abnormal lack of saturation. Roughness of surface is suggested as another possible cause. At sharp edges and points there will be excessively high fields, which may produce a large Schottky effect even at moderate voltages, thus causing an apparent deviation from the Schottky law.

Recently in the course of observation of pulse effects (cf. § 5·2) Sproull (1945) found that the voltage-dependence of the saturation current is greatest for the initial current when the anode voltage is first applied. After current has been drawn for a small fraction of a second and consequently decreased, the deviation from the Schottky effect is found to be smaller, although the slope of the curve of log i against $\sqrt{}$(voltage) is still of the order of twice the calculated value. This is attributed to a decrease in the variation of the work function over the surface. As current is drawn the work function increases owing to the electrolytic removal of Ba^{++} ions from the surface, but this effect will be greater for patches of lower work function. Thus, the variation in work function over the surface becomes less extreme and the consequent excessive lack of saturation is less pronounced.

Heinze (1933) working at temperatures up to 1150° K. and Huber (1941) in the range 550°–650° K. found that their results were best represented by an empirical relation of the form

$$i_s = i_s^0 (V_A + V_{CA} + 1)^\beta, \qquad \ldots\ldots (5)$$

where V_{CA} = contact potential between cathode and anode.

β = constant for each particular cathode.

According to results of Krieg, quoted by Hermann and Wagener (1944, p. 151), the value of β, obtained from a double logarithmic plot, is greater for rough (sprayed) than for smooth (cataphoretically coated) cathode surfaces. Other empirical equations have been suggested by Gill (1937) and Wheatcroft (1940).

§ 3. PROCESSES DETERMINING THE EMISSION

3.1. *Introduction*

It is generally accepted that the 'active material' in oxide cathodes is the free alkaline earth metal, but there are divergent views regarding the mechanism of its formation, its distribution throughout the cathode, and the mechanism by which it enhances the electron emission from the oxide. In the following paragraphs the basic processes which determine the emission will be discussed. We are concerned with

(*a*) processes directly affecting the passage of the current from the core, through the coating and into the vacuum, i.e. conduction and contact phenomena;

(*b*) processes determining the distribution of the free metal, i.e. diffusion and evaporation.

3.2. *Conduction*

3.21. *Experimental data*

In view of the relative instability of the pure alkaline earth oxides in the ordinary atmosphere and the difficulties of accurate measurements in vacuum, it is not surprising that there is relatively little available information on their electrical conductivity, and that what there is tends to be conflicting. The experimental conditions employed in the different investigations have also varied considerably. Horton (1906) and Spanner (1924) measured the conductivity in air on oxide pellets formed by pressing between massive platinum electrodes in the furnace. Meyer and Schmidt (1932) and Albricht (1931) carried out conductivity measurements coupled with emission current determinations *in vacuo* on pellets held between nickel and platinum electrodes respectively; the pellets were formed initially from the carbonates which were decomposed by high temperature heat *in situ*. Blewett (1939 a) also used pellets for his determinations *in vacuo*.

Clausing and Ludwig (1933), Becker and Sears (1931), Kroczek and Lübke (1930), Nisibori and Kawamura (1940) confined their attentions to indirectly heated oxide coated cathodes prepared normally from the carbonates, but arranged to carry metal probes embedded in the coating. Conduction between the metal core and the metal probe in the coating was then measured at various temperatures, in parallel with the electron emission to an external anode. Reimann and Murgoci (1930) and Reimann and Treloar (1931) used two oxide-coated filaments twisted together and recoated, thus allowing simultaneous measurement of the conduction current between the core filaments and of the emission from the composite cathode to an external anode. (It should be noted here that, as pointed out by Blewett (1939 a), the value and the interpretation to be placed on these results employing probes are obscured by the possible presence of steep thermal gradients.)

Reimann and co-workers obtained current voltage curves of the type shown in figure 4, indicating non-ohmic conduction, especially at high temperatures. Nisibori and Kawamura (1940), however, obtained curves which indicated an ohmic resistance. Wright (1947)* on the basis of his more recent investigation of the conduction current in oxide cathodes suggests that the non-ohmic conduction which he also observed is due to a barrier layer between core and coating (cf. § 3.32), and that where ohmic conduction is observed this barrier layer must be absent or very thin.

All workers in this field are agreed that the temperature dependence of the conductivity (σ) is given by a relationship of the form

$$\sigma = \alpha \exp[-\beta/T] \qquad \ldots\ldots(6)$$

or possibly

$$\sigma = \alpha \exp[-\beta/T] - \gamma \exp[\delta/T], \qquad \ldots\ldots(7)$$

where T is the temperature and α, β, γ, δ are constants for the particular oxides. Smekal (1928) considers the relationship given by equation (7) to be typical for all dielectrics; this equation is applicable to Horton's data for CaO–PbO (Horton, 1906) and to those of Phipps and Leslie for solid NaCl (1928) and was also shown by Becker and Sears (1931) to apply to the conductivity of BaO–SrO

* The authors are indebted to Mr. Wright for a preview of his paper at the proof stage.

in vacuo. The general exponential relation of the type of equation (6) was found by Spanner (1924) for alkaline earth oxides in air, and this was confirmed *in vacuo* by other investigators. It is, however, clear from figure 5, which shows

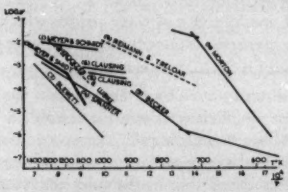


Figure 5. Electrical conductivity of BaO and BaO–SrO. (1) Meyer and Schmidt : activated BaO. (2) Meyer and Schmidt : unactivated BaO. (3) Kroczek and Lübke : conductivity of oxide coating on cathode (line given here is drawn through a number of widely scattered points obtained from measurements on different cathodes). (4) Spanner : conductivity in air of BaO in bulk. (5) Clausing : activated BaO on cathode (the temperature range over which these measurements and those of curve (6) were made is not given in the reference). (6) Clausing : conductivity of BaO on cathode after running for 1000 hours. (7) Blewett (1939 a) : conductivity in vacuum of BaO in bulk, rather loosely packed. (8) Reimann and Treloar : conductivity of BaO + SrO coating on filament (thickness of oxide is not given in the reference, so a value of 0·5 mm. was assumed in plotting this line). (9) Becker and Sears : conductivity of BaO + SrO coating on cathode. (10) Horton : conductivity in air of BaO in bulk (Blewett, 1939 a, figure 5).

the straight line plots of $\log \sigma$ against $1/T$ obtained from the results of ten different investigations, that there is little agreement regarding the values of the constants in the equation.

Kroczek and Lübke (1930) using alternating current measured the heat required to restore the original temperature after applying the anode voltage (cf. § 2.21, equations (2 *a*), (2 *b*), (3)). They found that the resistance of the mixed oxide is independent of the proportion of SrO present, but is determined solely by the barium content, and that the resistance increases with particle size and with the coating thickness. They state that the absolute value of the specific resistance is strongly dependent on the degree of activation.

The dependence of the conductivity on the frequency of the applied E.M.F. was examined by Williams and Huxford (1930) on indirectly heated cathodes. The conductivity decreased by about 40% when the frequency was altered from 15 000 to 10 000 c.p.s., and the decrease started at a lower frequency the higher the applied potential. On the other hand, Meyer and Schmidt (1932), using alternating current and oxides pressed between nickel blocks, found no dependence on frequency. These latter workers observed an increased conductivity during activation both with A.C. and D.C. potentials, but the resistance was much lower when D.C. was used. This was attributed to activation by electrolysis, the formation of barium sub-oxide (Ba_2O) of high electronic conductivity being postulated, but it has since been stated by Schriel (1937) that there is no evidence for the formation of this compound.

A parallel temperature dependence of the conductivity and the emission was first observed by Albricht (1931), and later by Reimann and Treloar (1931) and Nisibori and Kawamura (1940) at various temperatures. Nisibori and Kawamura obtained a linear relation between emission and conductivity during activation.

3.22. *Theory of the conduction process*

Two facts emerge from the diversity of experimental results regarding the conduction of alkaline earth oxides:—

(i) The electrical conductivity of these oxides which are insulators at room temperature increases exponentially with rising temperature.

(ii) The conductivity is ' structure sensitive '.

It was also pointed out in 1928 by v. Hevesy that the temperature coefficient of the resistance is not a sufficient criterion to establish the mechanism of conduction in a solid. Most, though not all, electrolytic conductors show a strong increase of conductivity with temperature, but so do some electronic conductors.

There is a deal of evidence that in the fully activated oxide a small fraction of the current is carried by alkaline earth metal ions (Becker, 1929, 1931; Berdennikowa, 1932), and even smaller proportions by oxygen ions (Spanner, 1924; Rothe, 1926; Espe, 1927; Vogt, 1931; Isensee, 1937). The actual proportion appears to vary for different oxide specimens.

Fowler and Wilson (1932) showed that the observed behaviour of oxide coated cathodes could be satisfactorily explained on the supposition that the conduction is mainly electronic at high temperatures, but mainly electrolytic at low temperatures. In particular, this conception, which was based on the then new theory of semiconductors, satisfactorily explained the results obtained by Reimann and Treloar (1931) which had previously been used as a main support for the theory of electrolytic conduction.

Although Fowler and Wilson's interpretation is partly based on a theory of rectification which has since been questioned by several authors (Schottky and Hartmann, 1933; Schottky and Spenke, 1939; Mott, 1939), there is no reason to doubt the basic conception of the alkaline earth oxide as an electronic semiconductor. It is now generally accepted that, although electrolysis plays an important part during the activation of the cathode (cf. §2.2), the greater part of the current in the fully activated coating is carried by electrons. It is also agreed that the conduction is caused, or at least enhanced by the presence of free alkaline earth metal. There are, however, two main views regarding the processes involved in the electronic conduction. One school of thought bases its conclusions on Wilson's theory of semiconductors (Wilson, 1931 a and b), in which the process of conduction is regarded as the thermal excitation of electrons from localized ' impurity ' levels in the lattice to the conduction band. A more exact theory on the same lines has been developed by Bronstein (1932). The other school of thought, represented by de Boer and his collaborators (de Boer, 1935 a and b; de Boer and van Geel, 1935), regards the process of conduction as one of thermal ionization of internally adsorbed atoms. It will be seen later that these two views are not as has been suggested (Fan, 1943), entirely equivalent and that they lead to different results.

Wilson first put forward his theory of semiconductors in 1931 (Wilson, 1931 a). It is based on the concepts of quantum mechanics, in which all electrons are considered to be free to move about in the periodic field of the lattice. In such a field, the energy levels are broken up into bands of allowed energy, separated by forbidden regions. If the highest band of allowed states, containing any

electrons at low temperatures, is only partly full of electrons, conduction can take place at all temperatures and the substance is a metal. This will be the case either if there are only enough valency-electrons to fill half the states in a single band or if a number of consecutive bands overlap. If, on the other hand, the electrons just fill an energy-band at absolute zero, no movement of electrons can take place under the influence of a small field because of the Pauli exclusion principle. Such a substance will be an insulator. With rising temperature, electrons may be excited to higher energy-bands and conduction then becomes possible. We thus have a model of an electronic semiconductor, in which the conductivity increases with rising temperature. At higher temperatures, however, the increasing thermal agitation of the ions of the lattice will come into play and the temperature coefficient of the resistance may finally become positive, as for metals.

Gudden (1931) first expressed doubt whether 'intrinsic' semiconductors, for which the Wilson model was derived, do in fact exist. He suggested that, in most known semiconductors, the conductivity is due to some impurity. This may be either a foreign atom or an excess amount of one of the component elements of the compound. For instance, the conductivity of cuprous oxide depends on the amount of excess oxygen present (Jusé and Kurtschatow, 1932, and 1933; Gundermann and Wagner, 1937). Wilson (1931 b) accordingly modified his theory to account for such cases. In the case of a semiconductor containing an 'impurity', the electrons of the impurity-atoms are not free to move through the lattice, unless their energy is the same as that of an empty band of the pure substance. The impurity electrons are thus bound to localized levels (figure 6). Wilson considers the problem of conduction in a semiconductor as one of equilibrium between 'bound' electrons, i.e. electrons in a full band or at a localized impurity-level, where they cannot contribute to the conduction, and 'free' electrons, i.e. electrons in the 'conduction band'.

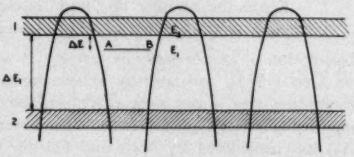


Figure 6. Energy levels in impurity semiconductor (after Mott and Gurney, 1940).
Band 1: Conduction band, empty.
Band 2: Highest full band.
Level A, B: Localized discrete level of an impurity-atom.

If the free electrons in band 1 are supplied primarily by the localized impurities, band 2 can be ignored. The dissociative equilibrium is then:

Bound electron $\rightleftharpoons$ free electron + bound hole.

Quantum statistical treatment of this problem (cf. Fowler, 1936, p. 397; Mott and Gurney, 1940, p. 158) leads to

$$N_1/v = (N_l/v)^{1/2}\varpi_1^{\,2}(2\pi mkT/h^2)^{3/4}\exp[-\Delta E/2kT], \qquad \ldots\ldots(8)$$

where N_1 = number of free electrons in a volume V (supplied from impurity levels).

N_l = number of impurity levels.

m = effective mass of electrons in states near the limits of band 1.

ΔE = energy difference between the bottom of band 1 and the impurity level A B.

w_1 = statistical weight of states of band 1.

h = Planck's constant.

The conductivity (σ) is given by

$$\sigma = (evN_1)/V,$$

where v is the mobility of the free electrons and is a slowly varying function of the temperature. Thus, the temperature-dependence of the conductivity is given by

$$\sigma = \sigma_0 . \exp[-\Delta E/2kT], \qquad \ldots\ldots(9)$$

where σ_0 varies slowly with the temperature and is proportional to the square-root of the impurity concentration. This theory has been generalized in a more systematic way by Fowler (1936).

The above derivation is based on these two assumptions:

(i) The number of 'free' electrons is much smaller than the number of impurity-electrons available.

(ii) All impurity-levels are occupied by electrons at low temperatures, and hence the number of 'free' electrons at any higher temperature is equal to the number of holes in the impurity levels.

Assumption (i) cannot easily be tested. It was first pointed out by de Boer and van Geel (1935), and later by Wilson himself (Wilson, 1939) that the second assumption also is not necessarily justified. The case where the number of impurity-levels (N_1) is much greater than the number of 'free' electrons available (N_1) was considered by Mott and Gurney (1940, p. 159). If assumption (i) still holds then for $N_1 \ll N_1 - N$:

$$N_1/V \propto \{N/(N_1 - N)\} \exp[-\Delta E/kT], \qquad \ldots\ldots(10a)$$

where N = total number of electrons, of which N_1 are excited (i.e. 'free') at the temperature T; and $N_1 \gg N_1$.

But for $N_1 \gg N_1 - N$:

$$N_1/V \propto (N_1/V)^{1/2} \exp[-\Delta E/2kT]. \qquad \ldots\ldots(10b)$$

There is a critical temperature (T_c) given by

$$2(N_1 - N)^2/V(2N - N_1) = (2\pi mkT_c/h^2)^{3/2} \exp[-\Delta E/kT_c], \qquad \ldots\ldots(11)$$

where the slope of the curve of log σ against $1/T$ changes from $-\Delta E/k$ to $-\Delta E/2k$ (figure 7).

Mott and Gurney also consider the possibility of an 'intrinsic' semiconductor containing impurity-centres as well. In this case, the conductivity will be given by the sum of two terms

$$\sigma = A_1 . \exp[-\Delta E_1/kT] + A_2 . \exp[-\Delta E/kT], \qquad \ldots\ldots(12)$$

where ΔE_1 = activation energy of intrinsic conduction (figure 6).

ΔE = activation energy of conduction due to impurity.

The term due to intrinsic conduction has a large activation energy and a

large temperature coefficient, while the term due to the impurity has a smaller activation energy and a much smaller coefficient (figure 8).

These considerations apply to electronic semiconductors in general. Experimental work relating to the theory has been carried out mainly on cuprous oxide (v. Hevesy, 1928; Jusé and Kurtschatow, 1932, 1933; Dünwald and Wagner, 1933; Brattain, 1933; Le Blanc, Sachse and Schröpel, 1933; Gundermann and Wagner, 1937; Gundermann, Hauffe and Wagner, 1937; Joffe and Joffe, 1940; Feldman, 1943), aluminium oxide (Podszus, 1933; Hartmann, 1936), magnesium oxide (Shul'man, 1940; Rousseau, 1934), zinc oxide (v. Baumbach and Wagner, 1933; Fritsch, 1935; Meyer and Neldel, 1937), nickel oxide (v. Baumbach and Wagner, 1934), and molybdenum trioxide (Ars'eneva and Kurtschatow, 1933). Fowler and Wilson (1932) first pointed out that the semiconductor theory also offers a basis for explanation of the main phenomena exhibited by the alkaline earth oxides, if these are assumed to contain excess alkaline earth metal atoms acting as the 'impurity'. There is ample evidence for the presence of an excess of the metallic component in the fully activated oxide (Berdennikowa, 1932; Fritsch, 1935; Prescott and Morrison, 1938).

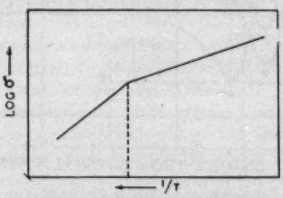


Figure 7. Conductivity-temperature curve for a semiconductor with more impurity-centres than electrons (Mott and Gurney, 1940, figure 60 (b)).

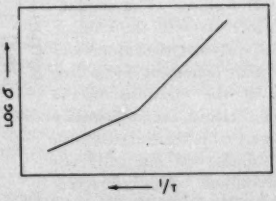


Figure 8. Conductivity-temperature curve for an intrinsic semiconductor with impurity-centres (Mott and Gurney, 1940, figure 60 (a)).

Reimann (1934), Blewett (1939 a), Uehara and Takahari (1940), and Nisibori and Kawamura (1940) all base their interpretation of conduction and emission phenomena on this concept (§ 4). It is dependent on the assumption that the excess metal (e.g. Ba) is actually present within the oxide-lattice, either in interstitial positions or at normal lattice points. In the latter case, the excess over the stoichiometric proportions is caused by the absence of some oxygen-ions from their normal lattice-positions. Various aspects of these two types of impurity-centres are discussed by Mott and Gurney (1940). According to these authors, the latter type of irregularity leads to the formation of the so-called *F*-centres, due to an electron being trapped at a point where a negative ion is missing. They also point out that no substance is known in which these give rise to electronic conductivity at temperatures at which the electrolytic conductivity is negligible. As yet the alkaline earth oxides have not been shown to have the characteristic absorption bands that are associated with *F*-centres.

It was first suggested by de Boer and van Geel (1935), that Wilson's assumption

of a homogeneous equilibrium between 'bound' and 'free' electrons and 'holes' is not justified in the case of a real impurity-semiconductor. They consider that the built-in impurity atoms are to be regarded as adsorbed on inner surfaces (so-called ' Lockerstellen ') and do not strictly belong to the lattice. Conduction is caused by the thermal ionization of the internally adsorbed atoms and similar considerations apply as for the problem of emission from atoms adsorbed on the surface (cf. § 4); but instead of being emitted into the vacuum the electrons are emitted into one of the permitted bands of the lattice (figure 9). These electrons move through the lattice in the empty permitted band and thus contribute to the conduction, until they are trapped by an ion on an internal surface. The conductivity increases with the number of internally adsorbed atoms and also with increasing mean free path of the electrons in the lattice. Hence, the conductivity may vary more than linearly with the number of internally adsorbed atoms. This is in agreement with experimental results for some compounds (Joffe, 1933).

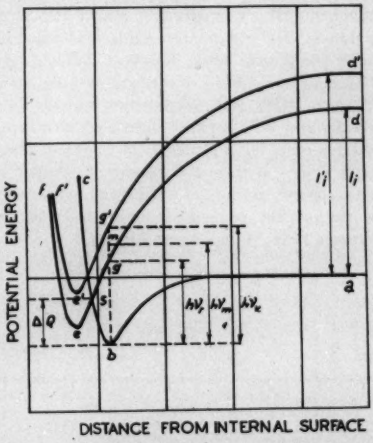


Figure 9. Potential curves for the ionization of an internally adsorbed atom (de Boer, 1935 a, figure 126).
abc = potential energy of an internally adsorbed atom, *a* representing the free atom.
def to *d'e'f'* = potential energy band for an internally adsorbed ion.
I = ionization energy of the free atom, between I_i and I'_i.
$h\nu_m$ = ionization energy of the adsorbed atom between $h\nu_I$ and $h\nu_\varphi$.

de Boer and van Geel contend that the ionization-energy decreases and may eventually even become zero, with increasing number of adsorbed atoms. This view is also held by Möglich (1938), who put forward a very similar theory for the conduction of compounds containing an excess of one component. There is considerable experimental evidence that the activation-energy of the conductivity of such compounds does in fact depend on the impurity-concentration (Jusé and Kurtschatow, 1933; Fritsch, 1935; Hartmann, 1936; Gundermann and Wagner, 1937; Meyer and Nedel, 1937).

This theory was developed mainly on the basis of conductivity data on cuprous oxide, nickel oxide and zinc oxide, and was applied by de Boer (1935 a, 1937) to dielectrics containing adsorbed atoms generally, such as alkali-halides with adsorbed alkali metal atoms, and oxide cathodes. An up-to-date qualitative review of the data available on semiconductors is given by Seitz (1945) and Maurer (1945).

3.3. *Contact phenomena and the core–coating interface*

3.31. *Contact phenomena*

In its passage through the oxide cathode the electron encounters two changes of medium, each involving a discontinuity in the potential energy. The processes taking place at the boundary between a metal and a semiconductor, such as the core–coating interface, have been discussed by various authors (Fowler, 1933; Mönch, 1933; Schottky and Hartmann, 1933; Mott, 1938, 1939; Schottky and Spenke, 1939; Davydov, 1939; Pekar, 1940, 1941; Joffe and Joffe, 1941; Gisolf, 1942; Fan, 1942; Lyashenko, 1942; Joffe, 1946).

A simple exposition of the modern theory has been given by Mott and Gurney (1940, p. 174). Figure 10 shows the relative position of energy-levels in the metal and an impurity-semiconductor.

Figure 10. Contact between a metal and a semiconductor.
– – – – – denotes position of impurity levels (Mott and Gurney, 1940, figure 69).

It is assumed that the impurity levels in the semiconductor are above the highest occupied band in the metal, as is shown in figure 10 (a), where the two substances are not in contact. When the metal and the semiconductor are brought into contact, the impurity centres in the layer of the semiconductor near to the metal will lose electrons to the metal. A positive space charge will be set up in the semiconductor which induces a negative charge on the surface of the metal. A field is set up which will tend to draw electrons away from the metal. Thus the energy of electrons in the lowest state of the conduction band in the semiconductor will be lowered, until the impurity levels are below the level of the Fermi-distribution (figure 10(b)). The breadth of the potential barrier (AB) is estimated to be $\sim 3 \times 10^{-7}$ cm., indicating that only in a layer ~ 20 atoms thick will all the impurity centres lose their electrons. It can be shown (Mott and Gurney, 1940, p. 175) that, in equilibrium, the lowest state in the conduction-band will have an energy ΔU above the surface of the Fermi-distribution in the metal, where ΔU is given by

$$\Delta U = \Delta E/2 - \boldsymbol{k}T \,.\, \ln \frac{N_1^{1/2}}{2(2\pi \boldsymbol{mk}T/\boldsymbol{h}^2)^{3/2}}. \qquad \ldots\ldots(13)$$

To a first approximation, at low temperatures

$$\Delta U = \Delta E/2.$$

This result has been used in the interpretation of emission phenomena on the basis of Wilson and Fowler's semiconductor theory (Reimann, 1934; Blewett, 1939a; Uehara and Takahari, 1940), but no experimental data are available regarding the energy-values involved in any real case. It is outside the scope of this review to attempt to correlate the work done in the field of contact phenomena for metals and semiconductors with the problem of the emission mechanism; however, Joffe (1945) has recently reviewed the theories of contact phenomena mainly with a view to their application to solid rectifiers. A study of the considerations put forward with regard to contact problems should provide new ways of tackling the emission question, both theoretically and experimentally.

3.32. *Rôle of the core metal and the interface*

It seems likely that, as Reimann suggests (Reimann, 1934), the effect of the core materials on the emission is an indirect one. Reducing properties of the core will promote the formation of free alkaline earth metal in the oxide and thus enhance the conductivity of the coating. An increase in the barium concentration will cause a rise in the ' impurity level ' and therefore also a decrease in the energy difference between the Fermi-level in the metal and the conduction-band, thus lowering the total work function (§4.2). There is no evidence for or against this hypothesis. There is, however, ample evidence that easily oxidizable components in the core do in fact enhance the emission (Lowry, 1930; Gehrts, 1930; Benjamin, 1935). The effects of Ti, Mg and Ni have already been referred to in §2.1. Liebold (1941) has attempted to correlate the work functions of oxide coatings on various core metals with the heats of formation of the core-metal oxides which he found were formed in the reduction process. The oxide coating (a barium, strontium, calcium oxide mixture of unstated composition) was thermally activated at the optimum temperature (§2.11), and given a current activation treatment at a lower temperature, the saturated emission was measured at a temperature where it was 10^{-5} to 3×10^{-4} amp/cm², and the values extrapolated to zero field were used in equation (1) to calculate the work functions (at the temperatures used) with the assumption that $A = 120$ amp. cm^{-2} deg^{-2}. The results show no direct correlation (figure 11), and to explain this Liebold postulates a balancing process between the reducing action of the core (assessed by the heat of formation of the core oxide) and the stability at high temperature and low oxygen pressures of the intermediate layer formed. Microchemical analysis has shown the presence in these intermediate layers with various core metals of alkaline earth platinates (Simon, 1928), tungstates (Liebold, 1941; Rooksby and Steward, 1946), molybdates, chromates etc. (Liebold, 1941). Rooksby and Steward, on the basis of x-ray diffraction results, have intimated the presence of tungstates of the type M_3WO_6 where M can be Ba and/or Sr and/or Ca. In the case of nickel cores containing small percentages of Al, Si or Ti and coated with BaO–SrO, Rooksby (1947) found that the interlayer contained barium aluminate, ortho-silicate or ortho-titanate respectively. (See also §5.2—Fineman and Eisenstein, 1946.)

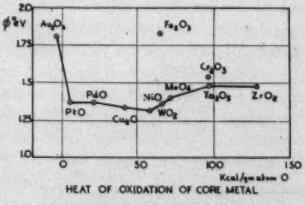


Figure 11. Relation between work functions for a triple oxide on various core metals and heat of oxidation of core metal to the oxide (indicated in brackets) considered to be formed (Liebold, 1941). Sprayed cathodes, optimum activation.

It is also conceivable that the composition of the oxide-coating may affect the potential barrier at the core-coating interface. The only information available regarding the effect of coating composition on the emission relates to variations in the ratio between the components of mixed alkaline earth oxides—BaO–SrO on platinum (Simon, 1927, 1928), on nickel (Benjamin and Rooksby, 1933, and

quoted by Reimann, 1934; Huber, 1941; Veenemans, 1943), CaO–SrO and CaO–BaO on nickel (Huber, 1941), and BaO–SrO with MgO on nickel (Hermann, 1938). Figure 12 shows the variation of work function with composition calculated by Huber from his results in the manner used by Liebold described above. For BaO–SrO Huber found maximum emission at 40–50 mol % SrO, compared with the value 60–70 mol % found by the other investigators. It is not possible at this stage to ascribe these variations with composition entirely to changes in the conditions at the oxide surface or the core–coating interface, or in the properties of the oxide itself. It has been adequately demonstrated by electron and x-ray diffraction measurements (Burgers, 1933; Gärtner, 1935; Darbyshire, 1938; Huber and Wagener, 1942; Eisenstein, 1946b) that immediately after the normal temperature treatment and in their preparation the surfaces of BaO–SrO cathodes consist essentially of SrO—to a depth of 300–2000 atom layers according to Huber and Wagener; these authors found a similar behaviour with BaO–CaO and SrO–CaO cathodes, in each case the more volatile component being absent in the surface layers. On this basis de Boer (1937) and Huber and Wagner (1942) ascribed the observed variation in emission for the oxide pairs to changes in the semiconductor properties, i.e. in the 'inner' work function (see § 4.2), and disregarded changes in the nature of the core–coating interface.

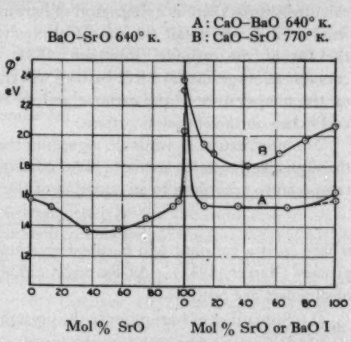


Figure 12. Variation work function ϕ_T (T=640 or 770° K.) with composition of alkaline earth oxide binary mixtures. Sprayed coatings on nickel activated at optimum temperatures. Means of 5–10 tubes (Huber, 1941).

In the absence of a clearer picture of the nature of, and processes occurring in, such intermediate layers at the high working temperatures, an evaluation of the rôle of contact phenomena in the emission process cannot be made.*

3.4. *Diffusion*

It is generally accepted that at any particular temperature there will be a net rate of diffusion of alkaline earth metal atoms towards the surface of the oxide,

* Wright (1947) has recently reported probe measurements on oxide cathodes with nickel bases containing magnesium or aluminium operated under D.C. conditions. His results provide evidence for, and are interpreted and discussed on the basis of, the presence of a potential barrier at the core–coating interface due to the formation of an interlayer which gives rise to contact phenomena either of the insulating (Mott, 1939) or semiconducting (Schottky, 1939 ; Mott and Gurney, 1940) type. With well activated and well aged coatings the height of this barrier is considered to determine the conduction as well as the emission current. The general problem in relation to rectification at metal–semiconductor contacts has recently been treated by Bardeen (1947) for the case where localized Tamm surface states, with energies in the 'forbidden' range between the filled and conduction bands, may exist at the surface of the semiconductor.

in order to replace those removed by evaporation or possibly by recombination with oxygen. Against this outward diffusion of barium (and/or strontium and/or calcium) atoms there is a migration of barium-ions towards the core, which manifests itself in the small amount of electrolytic conduction observed throughout the life of the cathode (Reimann, 1934; Blewett, 1939a). An equilibrium concentration-gradient of free barium will thus be maintained, which will depend on the temperature of the cathode and on the potential drop across the coating, and hence on the emission current.

No information is available regarding the rate of diffusion of the metal atoms through alkaline earth oxides. The diffusion coefficient (D) will depend on the temperature according to an equation of the form:

$$D = D_0 \exp[-E_d/\boldsymbol{R}T], \qquad \ldots\ldots(14)$$

where D_0 is a constant and E_d the activation-energy (in calories) of the diffusion process (Barrer, 1941). As Reimann (1934) points out, two types of diffusion mechanism are possible:

(1) Spreading of barium over the crystal surfaces.

(2) Diffusion of barium into the interior of the oxide crystals.

Reimann considers the first process the more likely, at least at low temperatures. This is in accordance with the picture of 'barium circulation' put forward by Reimann and Murgoci (1930), who supposed that barium ions are displaced from the surface to the inside of the coating through the crystal, whereas barium atoms migrate from the inside to the outside along the crystal boundaries (cf. § 4). de Boer (1935a, p. 352) also suggests that the barium atoms probably diffuse along the crystal boundaries.

It is generally supposed, that the activation-energy necessary for grain-boundary diffusion is less than for 'volume diffusion' through the lattice and that the latter process is more likely to come into play at higher temperatures (Barrer, 1941). Blewett (1939b) obtained a value of 17 kcal. (0·7 ev.) for the activation-energy of diffusion of Ba atoms through an oxide coating, from the temperature dependence of the recovery process after drawing emission. This compares with the value ~30 kcal. found for diffusion of lead atoms in lead chloride (Hevesy and Seith, 1929). No general expression is available regarding the temperature-dependence of the diffusion coefficient for grain boundary diffusion.

For diffusion of interstitial atoms or ions, an expression for the temperature dependence has been derived theoretically (Mott and Gurney, 1940, p. 33):

$$D = (\gamma BC)\nu a^2 \sqrt{(N_i'/N_a)} \cdot \exp[-(\tfrac{1}{2}W_0 - U_0)/\boldsymbol{k}T], \qquad \ldots\ldots(15)$$
$$= D_0 \exp[-A/T],$$

where γ, B and C are constants.

ν = vibration frequency of the interstitial atom or ion.

a = distance between adjacent interstitial positions.

N_i' = total number of possible interstitial positions.

N_a = total number of atoms (or ions) in the lattice.

U_0 = potential barrier between interstitial positions at absolute zero.

W_0 = work necessary to form a vacant lattice point at absolute zero.

According to Mott and Gurney (1940), D_0 can be expected to have values greater than unity for diffusion due to mobile lattice defects or to any mechanism characteristic of the ions in the lattice, but not due to impurity defects or to motion along cracks. For interstitial ions, these authors give an expression for the ratio between the ionic conductivity σ (in E.M.U.) and the diffusion coefficient

$$\sigma/D = n_a e^2/kT, \qquad \ldots\ldots(16)$$

n_a = number of ion-pairs per cm^3

For $n_a \sim 10^{22}$ and $T \simeq 600°$ K. this gives a value of $\sim 3 \times 10^4$ for all substances, if σ is measured in ohm^{-1} cm^{-1}.

Determination of the temperature dependence of the diffusion rate in conjunction with electrolytic conductivity measurements should thus provide evidence regarding the mechanism of the diffusion. Taking the total conductivity at 1225° K. as 7×10^{-3} ohm^{-1} cm^{-1} (Meyer and Schmidt, 1932) the ratio of electrolytic to total conductivity as 5×10^{-4} (Isensee, 1937), and $n_a = 2{\cdot}35 \times 10^{22}$ ions cm^3, the diffusion coefficient for Ba^{++} in BaO at 1225° K. would be $\sim 1 \times 10^{-10}$ cm^2 sec^{-1} according to equation (16). This with Blewett's value of 17 kcal. for the activation energy gives $D_0 \sim 10^{-7}$ cm^2 sec^{-1}, which is considerably smaller than values deduced for ionic diffusion in halides (see Barrer, 1941, p. 274).

3.5. *Evaporation*

No quantitative data are available regarding the rate of evaporation of free barium from the oxide. Presumably it is dependent on the amount of barium adsorbed on the surface. Such a dependence has been found in the case of atomic films on metals, such as barium or caesium adsorbed on tungsten. According to de Boer (1935 a, p. 82), the heat of evaporation of caesium atoms from a tungsten surface decreases proportionally as the degree of covering with caesium increases, i.e. the rate of evaporation increases exponentially. From the variation of the emission with the amount of barium deposited on the oxide surface, either by condensation from outside or by electrolysis, Becker and Sears (1931) concluded that there is an optimum degree of covering for maximum emission. If the surface covering is more than the optimum, glowing will cause an increase in emission due to evaporation (and back-diffusion) of barium until the optimum value is reached. The emission then decreases until the equilibrium state for the particular temperature is reached.

It was observed by Becker (1929) that large quantities of barium can be distilled from an oxide cathode without affecting its activity. It is concluded that, although the emission is extremely sensitive to small changes in the degree of surface-covering, the latter is not so sensitive to changes in the concentration of barium in the 'reservoir' within the coating. Maddock (1935) contends that the equilibrium covering of the surface with barium decreases with rising temperature. As Blewett (1939 a) points out, there is at present insufficient evidence regarding the very existence of a layer of barium on the surface of the oxide cathode. Electron diffraction experiments have so far failed to detect the presence of any such layer (Gärtner, 1935; Darbyshire, 1938; Huber and Wagener, 1942).

§ 4. GENERAL THEORIES OF EMISSION

4.1. *Introduction*

At the time of publication of Dushman's review (1930) there were two main points of view regarding the origin of the emission from oxide coated cathodes. On the one hand, Becker (1929) felt that he had shown conclusively that at least part of the free alkaline earth metal was located on the surface of the oxide, probably as a monatomic layer maintained by diffusion of the metal from the interior; he considered that the electrons were emitted from this surface layer and that the conduction through the oxide coating was largely by electrons and to a small extent by barium ions (cf. § 3.2). On the other hand, Lowry (1930) held that the electrons were emitted from the layer of alkaline earth metal adsorbed on the core, and hence diffused through interstices in the oxide coating into the vacuum; he claimed that this theory accounted for the decay in emission during life, de-activation by overheating and non-saturation. Reimann and Murgoci (1930) agreed with Lowry regarding the origin of the electrons, but maintained that they were transferred through the coating by emission from crystal to crystal until finally emitted at the outer surface, and that within each crystal the conduction was electrolytic. They suggested that the process of barium circulation was a balance between the diffusion of barium along the crystal boundaries towards the surface, and the migration of barium ions within the crystal towards the core. Some corroborating evidence for this was drawn from the earlier work by Davisson and Germer (1924). Dushman in his review drew no definite conclusions regarding these theories.

The theory that a monatomic layer of barium on the surface of the oxide is responsible for the enhanced emission of oxide cathodes was further developed by Becker (1931), and by Becker and Sears (1931) who proved conclusively that the emission is independent of the area of the core surface, but is extremely sensitive to variations in the amount of barium covering the oxide surface. Becker and Sears suggested that the thermionic electrons originate in the oxide just beneath the barium layer, that is, that the conduction electrons of the oxide itself are emitted, the work function of the oxide being lowered by the absorbed barium on the surface; at the same time the A-value in the emission formula is increased by the barium dispersed throughout the coating. This picture of the emission mechanism is still accepted today by the representatives of what we may call the 'Wilson school' of the theory of semiconductors. de Boer's concept of thermal ionization of adsorbed atoms leads to a different picture of the emission mechanism. As in the theory of conduction (cf. § 3.22) we have in the theory of emission these two main schools of thought. Uehara and Takahari (1940) have proposed a theory for the emission mechanism based on a zone model of the oxide which differs somewhat. These three concepts will be treated separately in the following sections, and in § 4.5 the attempts to reach a conclusion in favour of any particular mechanism will be discussed.

4.2. *The Wilson model*

The application to the alkaline earth oxides of the general theory of semiconductors put forward by Wilson (cf. § 3.21) has led to the conception that the emission from oxide cathodes is to be regarded as a passage of conduction electrons

from the core-metal through a potential barrier at the core-coating interface, through the coating, and finally into the vacuum, possibly through a second potential barrier associated with the surface polarization due to alkaline earth metal adsorbed on the surface. The picture of the emission mechanism put forward in Reimann's book (1934) is based on this view. Chiefly on Becker's evidence, Reimann considers it probable that in the fully activated cathode there is a monatomic film of barium adsorbed on the outer surface of the oxide which is maintained by the process of barium circulation suggested by Reimann and Murgoci. This process, however, now serves only to replace the barium removed from the surface by evaporation or recombination with oxygen. Electrolysis plays only a small part in the conduction through the coating (except during activation), the greater part of the current being carried by electrons. Activation consists essentially in the formation of free alkaline earth metal, initially by chemical reduction and subsequently also by electrolysis.

The free metal first appears on the core-coating interface and thence diffuses through the coating. In the equilibrium condition, there will be adsorbed barium (and/or strontium and/or calcium) on all the free faces of all the crystals of the coating and possibly also on the core-metal surface. The free alkaline earth metal in the coating constitutes the 'impurity' which causes the oxide to behave as an impurity-semiconductor.

The potential energy change suffered by the electrons carrying the current from the core, through the oxide and into the vacuum is represented by the continuous line in figure 13. The critical energy of the Fermi-distribution in

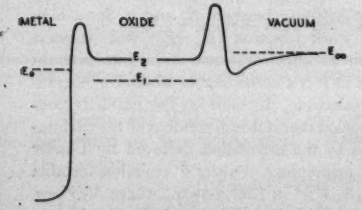


Figure 13. Schematic representation of energy changes in an oxide cathode (Reimann, 1934, figure 46).

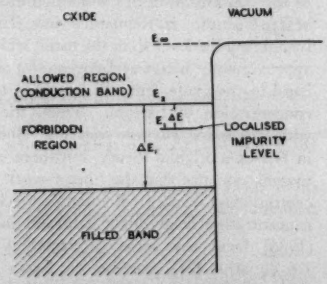


Figure 14. Energy level distribution for an impurity semiconductor (after Blewett, 1939 a, figure 6).

the core-metal is E_0. The energy level of the impurity electrons in the oxide lattice is E_1. The potential energy in the oxide (conduction level) is E_2. The potential energy in the vacuum is asymptotic to E_∞.

On the basis of the Wilson-Fowler theory, the distribution of the free conduction electrons (n_1) is Maxwellian since their number is small. Then the number n_2 of electrons emitted at $T°$ K. is

$$n_2 \propto n_1 \exp\left[-(E_\infty - E_2)/kT\right]. \qquad \ldots\ldots(17)$$

If the number of electrons thermally excited from the impurity levels is taken equal to n_1, then with equation (8) since $i_s^0 \propto n_2$,

$$i_s^0 \propto n_2 \propto \exp\left[-\frac{E_\infty - \frac{1}{2}(E_1+E_2)}{kT}\right]. \qquad \ldots\ldots(18)$$

(Compare equation (21).)

But it can be deduced statistically that

$$E_0 = (E_1+E_2)/2 \qquad \text{(cf. § 3.2 where } U = E_2 - E_0\text{)}.$$

Thus

$$i_s^0 \propto \exp\left[-(E_\infty - E_0)/kT\right]. \qquad \ldots\ldots(19)$$

Reimann points out that, although this expression does not directly involve the energy levels in the oxide (E_1 and E_2), these levels are in fact involved, since the potential energy of the metal must adjust itself so that $E_0 \sim \frac{1}{2}(E_1+E_2)$.

Reimann and Treloar (1931) had observed that the slope of the emission–$1/T$ Richardson line was only slightly greater than that of the straight line plot against $1/T$ of the 'apparent conductivity' of the oxide coating, the apparent conductivity being defined as the slope of the conduction current–voltage curve at the origin at any given temperature (figure 4). Reimann therefore concludes that the 'surface work function' $(E_\infty - E_2)$ is only a small fraction of the total work function $(E_\infty - E_0)$. The A-value in the Richardson equation is equal to $A_0(1-\bar{r})$ where A_0 is the same as for the uncoated metal and should be a universal constant, and $\bar{r}$ is the mean reflection coefficient for the surface of the oxide (see footnote on p. 299). The A-value is considered to be independent of the concentration of conduction electrons and therefore of the concentration of free barium. In Reimann's view, the barium concentration affects the emission by altering the level E_0 in the metal which, if there is no dissolved barium, will be approximately half-way between E_2 and the upper limit of the highest filled band in the oxide, but will be raised to half-way between E_1 and E_2 for a high concentration of barium. Thus, the work function $(E_\infty - E_0)$ will decrease with increasing barium content. Such an effect was not taken into account in Wilson's original theory, but there is in fact a considerable amount of experimental evidence that the 'inner work function' depends on the impurity concentration (cf. § 3.22). Reimann's conclusion that A is independent of the barium-concentration is in direct contradiction to the expression deduced by Fowler (1936) from the same fundamental principles. Fowler's emission-formula (cf. equation (21)) contains the factor $(N_1/V)^{1/2}$ in the A-term, where $N_1/V = n$ is the number of impurity-levels per unit volume. Thus the A-value is proportional to the square root of the number of impurity-atoms present.

Regarding the rôle of the core-material, Reimann accepts the evidence of Becker and Sears that the emission is not limited at the core. In his view, the core material probably affects the quantity of free barium in the coating, thus having an indirect effect on the emission.

In his general treatment of the problem Fowler (1936, § 11.62) takes the general equation obtained for pure metals:

$$i_s^0 = \varpi_r A T^2 \lambda, \qquad \ldots\ldots(20)$$

where ϖ_r = weight factor for the electron spin (= 2 for metals) and $2A = A_0$ and develops a statistical expression for λ which is then solved for particular

cases (cf. equations (10), (11)). For an oxide impurity semiconductor (cf. figure 14) at low temperatures when the average number of electrons in the conduction band $\bar{n}_1$ is very much greater than the average number of holes in the filled band,

$$i_s^0 = A_1 \bar{D} n_1^{1/2} T^{5/4} \exp\left[-\frac{E_\infty - \frac{1}{2}(E_2+E_1)}{kT}\right], \qquad \ldots\ldots(21)$$

where $A_1 = e.\varpi_r' \varpi_1^{-1/2}\left\{\frac{(2\pi m)^{1/2} k^{5/2}}{h^3}\right\}^{1/2}$, $\qquad \ldots\ldots(22)$

ϖ_r' = statistical weight factor for electron spin for the oxide, which is assumed = 2,

$\bar{D}$ = mean transmission coefficient for electrons at the oxide surface ($\bar{D} = 1 - r \simeq 1$),

and the other symbols have the meanings previously indicated.

For medium temperatures and $\bar{n}_1 \sim n_1$, the expression is

$$i_s^0 = 2e.\left(\frac{k}{2\pi m}\right)^{1/2}.T^{1/2}\exp\left[-\frac{(E_\infty - \Delta E)}{kT}\right], \qquad \ldots\ldots(23)$$

when $\bar{D}$ and ϖ_1 are put equal to 1 and $\varpi_r' = 2$.

In the case of an intrinsic semiconductor the equation is

$$i_s^0 = A_0 T^2 \exp[-(E_\infty - \tfrac{1}{2}\Delta E_1)/kT]. \qquad \ldots\ldots(24)$$

Equation (21) is identical in form with that derived independently by Schottky (1935). de Boer (1937) acknowledged the probable importance of a conductivity process and, on the basis of Schottky's model, derived an equation essentially similar to Fowler's equation (24):

$$i_s^0 = A_2 T^2 \exp[-(e\phi_s + \Delta E')/kT], \qquad \ldots\ldots(25)$$

where ϕ_s = 'surface' work function = $1/e(E_\infty - E_2)$ (figure 14),

$\Delta E'$ = energy interval between the filled impurity levels and the empty band.

$\Delta E'$ represents an 'inner' work function whose value is some 25% of that of the surface or outer work function.

Heinze and Wagener (1938: see also Hermann and Wagener, 1943, 1944) elaborated de Boer's treatment. With Maxwellian statistics, they obtained the equation

$$i_s^0 = n_{E_2} e(k/2\pi m)^{1/2} T^{1/2} \exp[-e\phi_s/kT], \qquad \ldots\ldots(26)$$

which is of similar form to the original empirical equation of Richardson (1914). The number of electrons in the free barium band with lower energy limit E_2, derived by application of Fermi-Dirac statistics to the impurity semiconductor problem, with $\varpi_r' = 2$ is

$$n_{E_2} = 2/h^3(2\pi mkT)^{3/2}\exp[e\phi_1/kT], \qquad \ldots\ldots(27)$$

which, with (26), gives

$$i_s^0 = A_0 T^2 \exp[-(\phi_1 - \phi_s)e/kT]; \qquad \ldots\ldots(28)$$

$e\phi_1$, the inner work function, is not equal to $(E_2 - E_1)$ in figure 14, but is given by

$$\phi_1 = (E_2 - E_1)/2e + \tfrac{1}{2}kT \ln\{(2\pi mkT)^{3/2}/h^3 n_1\} \qquad \ldots\ldots(29)$$
$$= E_2 - E_3.$$

This gives the temperature dependence of ϕ_1, allows E_2 to be calculated, and, when substituted in (28), leads to equation (21) (cf. figure 16 and equation (31)). (Equations (29) and (13) are identical if $e\phi_1 = \Delta U = E_2 - E_0$. Substitution for ϕ_1 in equation (28) leads to equation (8), indicating that $n_{E_2} = n_1$, the number of impurity atoms per unit volume.)

Blewett (1939 a) adopts the Fowler-Wilson picture, the major part of the free alkaline earth metal being considered as dispersed through the body of the oxide as 'atomic inclusions', supplying electrons both for conduction and thermionic emission. He makes use of equation (21), but in neglecting weight factors approximately equal to unity, he appears to have omitted the spin factor of 2. Values of the transmission coefficient $\overline{D}$ and the surface work function $(E_\infty - E_0)$ suggest the presence of a potential barrier at the oxide surface which may be due to a monatomic layer of free metal, but Blewett maintains that the experimental data are ' insufficient to enable us to speculate further regarding the nature of the oxide layer '.

4.3. *de Boer's model*

Blewett makes no reference to the entirely different concept of the emission mechanism put forward by de Boer (1935 a), who regards the emission as a process of thermal ionization of the alkaline earth metal atoms adsorbed on the oxide surface, the emitted electrons being supplemented by the ionization of internally adsorbed alkaline earth metal atoms. Thus, it is the valency electron of the adsorbed atom that is emitted.

Figure 15 shows the potential energy curves for the adsorption of an atom or an ion on a surface. For the free, non-adsorbed atom, the thermal energy of ionization is I, corresponding to the energy difference a–d. For the adsorbed atom, excited by thermal energy, there is a certain probability of transition from the lower curve to the higher one at the point of intersection S, so that the ionization energy for the adsorbed atom is ΔQ, corresponding to the energy difference b–s. It was first pointed out by de Boer and van Geel (1935), that the energy necessary for the thermal ionization of an adsorbed atom is smaller than that required for photo-ionization, which is represented by the energy difference b–g. This is a consequence of the Franck-Condon principle (Franck, 1930) which states that if a quantum of light is absorbed by an atom the transition from the atomic to ionic state has the greatest probability when the distance from the surface to the adsorbed atom or ion remains unchanged.

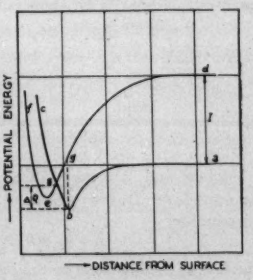


Figure 15. Potential curves for the adsorption of an atom (*abc*) or an ion (*def*) on a surface (de Boer, 1935 *a*, figure 107).

Similar considerations apply to the internally adsorbed atoms, which are responsible for the supplementing of electrons through the coating (§ 3.22). These atoms are supposed to be adsorbed on so-called ' Lockerstellen ', i.e.

'internal surfaces, places of disturbance, crystal imperfections, irregularities of the regular atoms or foreign atoms' (de Boer, 1935 a, p. 236).

With Maxwellian statistics, this conception leads to the emission-formula:

$$i_s = A_1 \exp\left[-e\phi_1/kT\right], \qquad \ldots\ldots (30)$$

where $e\phi_1 = \Delta Q$ = thermal ionization energy of the adsorbed atom.

A_1 = constant.

This equation differs from equation (1) only in the absence of the T^2 term, the emission constants will thus be slightly different from A and ϕ_0 in equation (1).

For the same dielectric and the same sort of adsorbed atoms, the value of A will depend on

(1) The number of adsorbed atoms per cm² of surface and hence also on the roughness of the surface.

(2) The number of times per second that an emitted electron can be supplemented and hence on the occupation of the lattice by internally adsorbed atoms (§ 3.12).

However, de Boer maintains that ϕ_1 will not, in the first instance, depend on the number of atoms adsorbed on the surface. So long as the portions of the surface covered are not large enough for regions to be formed where conduction electrons may be considered to exist, the work function ϕ_1 will be determined entirely by the activity of the most active spots on the surface. This statement has since been criticized by Fan (1943), who points out that the amount of adsorption must affect the difference between the energy levels inside and outside the oxide, on account of the double layer effect, and therefore the work function should vary with the degree of adsorption. Contrary to earlier schools of thought (e.g. Becker), de Boer visualized the oxide surface as consisting of isolated barium atoms rather than a complete monatomic film.

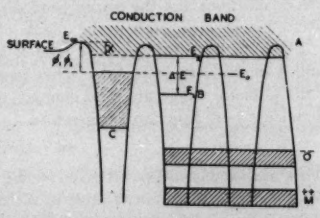


Figure 16. Energy diagram proposed by Uehara and Takahari (1940, figure 4).

A = lower limit of the conduction band in the oxide.
B = impurity level.
C = filled band of the valency electrons of the alkaline earth metal on the surface.
x = energy difference between vacuum and conduction level ($E_\infty - E_1$).
ΔE = energy difference between conduction level and impurity level ($E_2 - E_1$).
ϕ = thermionic work function.
ϕ_p = photoelectric work function.

The later German edition of de Boer's book (1937) contains some expansion of the original theory, particularly with regard to the conduction process (cf. equation (25)). The author emphasizes that, with the knowledge available at the time, it was not possible to decide definitely between the two opposed concepts of the emission mechanism.

4.4. *The model of Uehara and Takahari*

The model of the oxide cathode as a semiconductor with the emission caused directly by the impurity atoms has been criticized by Uehara and Takahari (1940).

They reject this theory on the ground that experimental results fail to bear out the conclusion that the thermionic work function is smaller than the photoelectric work function. Uehara and Takahari base their own theory on Huxford's result (1931) that the thermionic and photoelectric work functions are equal. They propose a zone-model (Uehara, 1933) for oxides, based on quantum mechanical theory of solids, taking into account the metallic atoms adsorbed on the surface. It is stipulated that the adsorbed metal is present as a two-dimensional layer, at least locally, even when the surface is not completely covered. The energy-diagram is shown in figure 16.

As before, the Fermi-surface in the oxide lies approximately halfway between A and B. The energy (E_0) of the Fermi-surface is given by

$$E_0=(E_2-E_1)/2-\boldsymbol{k}T\log\{B(\boldsymbol{k}T)^{-3/4}\}, \qquad \ldots\ldots(31)$$

which follows from (13) with $\Delta U=E_2-E_0$; the second term being negligible for $T\sim1000°$ K. and the number of electrons per unit volume in the B level $\sim10^{18}$ (Fröhlich, 1936, p. 226). Since the oxide and the metal on the surface are in equilibrium, the Fermi-surface in the adsorbed metal is the same. For an alkaline earth metal with a two-dimensional regular lattice, the S-band of the valency electrons and the empty band directly above it overlap, and the Fermi-surface lies approximately in the middle of the overlapping part.

Electrons from the impurity level B are thermally raised to the conduction band A of the semiconductor. They pass on their heat-energy to the electrons in the C band, which are then emitted over the potential barrier. Hence, the thermionic work function ϕ is given by

$$\boldsymbol{e}\phi=\Delta E/2+x. \qquad \ldots\ldots(32)$$

This is the same as the result given by the Fowler-Wilson theory, based on direct emission from the impurity level in the oxide. But in that case, if the photoelectrons are supposed to be of the same origin as the thermionic electrons, the photoelectric work function ϕ_p is given by

$$\boldsymbol{e}\phi_p=x+\Delta E=\boldsymbol{e}\phi+\Delta E/2. \qquad \ldots\ldots(33)$$

In the new concept, if the photoelectrons are assumed to come from the C band of the adsorbed metal atoms, the photoelectric work function is

$$\boldsymbol{e}\phi_p=x+\Delta E/2=\boldsymbol{e}\phi. \qquad \ldots\ldots(34)$$

Thus, this picture leads to equal photoelectric and thermionic work function, as required by the experimental result.

Uehara and Takahari's theory resembles de Boer's concept in that the metal atoms adsorbed on the surface are regarded as the origin of the emitted electrons. However, as Uehara and Takahari themselves point out, de Boer's theory does not require the assumption of a two-dimensional lattice for the adsorbed metal. In view of the fact that no barium has been detected on the surface of oxide cathodes by electron diffraction experiments (Darbyshire, 1938; Huber and Wagener, 1942), there is no evidence in favour of the model of Uehara and Takahari other than Huxford's isolated experimental results regarding the relationship between photoelectric and thermionic work function. Considering the difficulties in the interpretation of photoelectric data, this does not seem to

be sufficient basis for rejection of the comprehensive theories put forward by Fowler, Wilson, Reimann and Blewett on the one hand, and by de Boer and his collaborators on the other.

4.5. *General considerations*

An attempt was made by Heinze and Wagener (1936, 1938) to reach a conclusion in the controversy between the ' Wilson-model ' and de Boer's concept. They concluded from electron optical measurements and contact potential measurements (§ 2.1) that the surface-covering of the oxide with barium does not change appreciably during the activation. This is contrary to expectations on the basis of de Boer's theory. Heinze and Wagener attribute the increase in emission essentially to a decrease in the inner work function, that is to say, the energy required to raise the electrons to the conduction band. It has been previously noted that there is good evidence that this quantity depends on the concentration of the excess alkali metal within the coating (cf. § 3.22). Equation (29) shows the expected dependence of ϕ_i on n_i, and also on temperature. In the emission equation (21) the general constant A and not the total work function is dependent on n_i and T, which is contrary to experience. On this basis Heinze and Wagener accept equation (28) as their working formula from which to calculate the total work function (cf. § 3.32, figures 11 and 12). It is, however, pointed out that the derivation of equation (26), by Maxwellian statistics, and thus the validity of equation (28), is sound only when n_i is not so large that the critical temperature (Fowler, 1936, § 11) approaches the working temperature; the fact that ϕ increases only up to a certain value of n_0 and then decreases (Prescott and Morrison, 1938) may thus be explained.

Discussing the relationship between conductivity and emission characteristics of caesium oxide cathodes, Borzyak (1939) points out that the observed increase in the inner work function with decreasing impurity content does not necessarily imply any change in the ' impurity levels ', but may be due to the electrons of the highest filled band in the oxide coming into play to a greater extent as the impurity content decreases. This holds good only if the filled band does not lie at too great a depth.

Hermann and Wagener (1944, p. 193) have used Huber's total work functions (figure 14), obtained on the basis of equation (28) to estimate values of the outer work functions for the alkaline earth oxides on nickel. The values of the inner work functions were estimated very approximately (method not explicitly stated) from the data available, the starting point being the value $\phi_i = 0{\cdot}3$ ev. for BaO obtained by Clausing (1933). Their results are shown in figure 17, where M′/M″O represents for ϕ_i, excess metal M′ in oxide M″O, for ϕ_a, metal M′ adsorbed on oxide surface M″O. It should be noted that these values apply to the temperature region 650–750° K.

Morgulis (1941) has reviewed the thermionic, photoelectric and secondary emission characteristics of barium oxide, caesium oxide, antimony-caesium cathodes, aluminium oxide–caesium oxide (the ' Malter-cathode ') and cuprous oxide. The behaviour of these ' active semiconducting cathodes ' is considered essentially from the point of view of the Wilson-Fowler theory. Morgulis contends that the work function of a metal film adsorbed on a semiconductor is

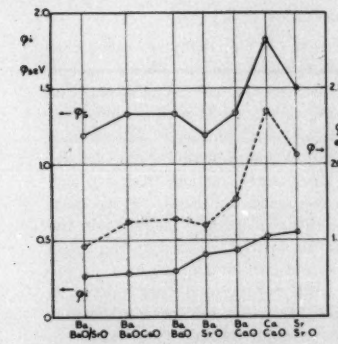


Figure 17. Inner (ϕ_i) and outer (ϕ_o) work functions (at 650–750° K.) estimated approximately from total work functions (ϕ_T) for various impurity metal/oxide systems (after Hermann and Wagener, 1944, p. 193). Sprayed cathodes on Ni base, activated at optimum temperatures.

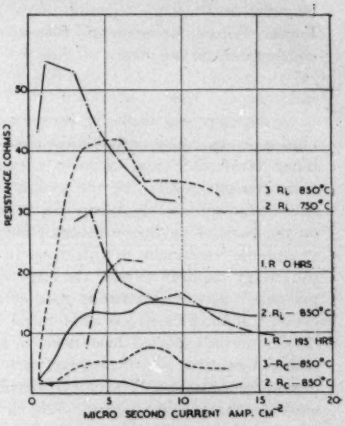


Figure 18. Variation of coating resistance with pulse current density.

1. Total resistance (estimated from pulse temperature rise) after 0 and 195 hours life ; BaO–SrO on 'A' Ni at 850° c. (Buck).
2. Interface (R_i at 750 and 850° c.) and coating (R_c at 850° C.) resistances by Ni probes ; BaO–SrO on electrolytic Ni.
3. R_i and R_c at 850° c. by Pt probes ; BaO–SrO on Cr-plated Ni (Fineman and Eisenstein).

smaller than for the same metal film on a metal base. Hence, the high emission of these compounds is attributed to one of two causes:

(i) a small work function of the pure oxide itself,

(ii) a large dipole moment of the metal film when adsorbed on the surface of a semiconductor.

Morgulis suggests that, in relation to oxide cathodes, full use has not yet been made of the methods used in the study of semiconductors, and that X-ray and electron-optical methods could be used to obtain much more information about the structure and surface properties of these systems.

§ 5. OTHER PHENOMENA OCCURRING WITH OXIDE COATED CATHODES

The main concern of this review has been the thermionic emission from oxide cathodes under 'normal' operating conditions. Attention has already been drawn to certain other properties of oxide cathode systems which may have an important bearing on the emission mechanism (§ 3), and to recent associated work on photoelectric emission (§ 4). The semiconducting properties and contact phenomena have been fully treated, and it is shown that material advances have been made during the last decade.

In recent years, however, some new and interesting properties of these systems have been described. It is only to be expected that, under the impetus given by war-time research and in view of the different requirements demanded from oxide cathodes under the new conditions, the number of these properties or at any rate information on existing ones, will increase rapidly. Any attempt to fit the new phenomena into a general picture of the emission mechanism must therefore be considered premature at the present time; this has already been indicated in the field of contact phenomena (§ 3.4) where further work must be awaited of the type initiated by the Joffes (1940, 1941, 1946) and by Fineman and Eisenstein (1946). Nevertheless, for the sake of completeness the results of these recent investigations will be briefly described.

5.1. *Luminescent properties*

Huber (1941) reported that oxide cathodes containing CaO emitted light (wavelength 5400–6600 A.) when emission currents exceeding 10^{-6} amp/cm^2 were drawn at anode voltages above 20 v., the light output increasing with emission current and anode voltage. The phenomenon was also observed with very pure CaO on very pure nickel bases, and was considered as further evidence for the presence of excess Ca in the CaO lattice, which then acted as a 'luminescent activator'.

5.2. *Thermionic emission under pulsed conditions*

In recent years it had been demonstrated that transient peak currents much higher than those obtainable under normal operating conditions can be drawn from an oxide cathode (with BaO–SrO coating unless otherwise stated) when a pulse of short duration is applied. Thus Schade (1943) obtained transient peaks of 25 amp/cm^2 in rectifiers with well activated cathodes, and Knight and Herbert (1946) mentioned the occurrence of transient peak currents of over 100 amp. in thyratrons with cathodes of unstated area. With microsecond pulses Andrianov (1944) reported values up to 30 amp/cm^2, and Coomes (1946) states that space-charge-limited values in excess of 100 amp/cm^2 have been obtained; according to Willshaw and co-workers (1946) current densities of 5–50 amp/cm^2 are readily obtainable in magnetrons, but secondary emission undoubtedly contributes to the high values here (see also § 5.3). These emission values compare with the maximum figure of the order of 0·5–1 amp/cm^2 for an oxide cathode under D.C. conditions at equivalent operating temperatures. Andrianov (1945) and Coomes (1946) found for BaO–SrO cathodes that in the space-charge-limited region the current/voltage characteristics followed the normal Langmuir–Child 3/2 power law, but Boot and Randall (1946) stated that numerous deviations were found for voltages up to 200 kv.

A short-time decay in the saturated emission current is generally observed. This has been studied for BaO–SrO on nickel cathodes containing 2% silicon over the time range 0·2–300 μsec. by Sproull (1945) who demonstrated that

(*a*) Over a wide range of thickness of oxide coatings the initial and final currents are approximately constant, i.e. the rate of decay is independent of the thickness.

(*b*) The rate of decay increases with increasing anode voltage.

(*c*) For a fixed frequency of repetition, the rate of decay is independent of the duration of the applied pulse.

(*d*) For a fixed duration of pulse, the rate of decay is independent of the repetition frequency.

Sproull's interpretation of this initial decay phenomenon is based on the assumption that the enhanced emission is due to a layer of barium atoms at the emitting surface. When a current is drawn, Ba^{++} ions flow towards the core under the influence of the field, causing a decrease in the barium concentration on the surface, and hence a decrease in the current. This flow is eventually balanced by the outward diffusion of Ba atoms. He developed an expression for the rate of decay of the form:

$$\frac{1}{i}\frac{di}{dt} = \beta \log\left(\frac{i_0}{i}\right) - \left(\frac{\alpha L}{e n_s}\right) i, \qquad \ldots\ldots(35)$$

where i_0 is initial pulsed current in amp/cm^2.

i is pulsed current at time t in amp/cm^2.

n_s is number of Ba atoms per cm^2 surface at the degree of covering $\theta = 1$.

α, β are constants.

L is a slowly varying function of θ, constant over a range of θ producing a change by a factor of 10 in the emission current.

With suitably chosen α and β values Sproull obtained theoretical curves in very good agreement with those experimentally determined.

The factors affecting the life of oxide cathodes under pulsed conditions—with special reference to magnetron cathodes—were examined during the war by English workers and by an American Cathode Research Group. The results of the English work have been summarized in a general paper by Willshaw and others (1946) and the American results have been presented in some detail in papers by Eisenstein (1946 a, 1946 b), Coomes (1946), Finemann and Eisenstein (1946), in which suitable reference to unpublished work of others has been made. Coomes and Willshaw state that no direct correlation between pulsed and D.C. emission properties were found; according to Coomes good pulsed cathodes were usually good D.C. cathodes, but the reverse was not always true. Because of the different phenomena and limitations the following new parameters were defined for assessing comparative merits under pulsed conditions:

(*a*) *Sparking points*—the value of the pulsed current (I_s) at which emission is limited by the physical transfer of particles from cathode to anode or 'sparking' (Schade, 1943; Willshaw, 1946).

(*b*) *Maximum space-charge-limited current*—the emission current where the $I/V^{3/2}$ curve deviates from a straight line.

(*c*) *Comparative temperature rise*—accompanying pulse emission at given current levels and duty cycles.

I_s was usually obtained at a point still on the Langmuir-Child characteristic line, when it represented the highest current that could be drawn from the cathode. During life I_s first increased, then remained approximately constant at a flat maximum, then decreased sharply; its value and variation depended on the ageing conditions, on the ageing temperature (a rapid decrease occurred above

800° c.), and on the nature of the nickel base. Values of I_s in excess of 100 amp/cm² were obtainable. The pulse-temperature rise, which was 5–60° c. for normal BaO–SrO at 800° c. with 1 microsecond pulse of pulse repetition frequency 1000 per sec., also varied during life; the minimum value occurred approximately at the point where I_s attained a maximum. Variations in pulsed emission occurring during life are also referred to by Pomerantz (1946 a, 1946 b). Willshaw (1946) reported that the primary emission in magnetrons decays exponentially during life, the time required to reach half the initial emission value decreasing rapidly with increasing cathode temperature. Willshaw also stated that the sparking phenomena, interpreted as a flash arc or voltage breakdown in the cathode, occur in all magnetrons at high current densities and voltages greater than 20 kv. during ageing. This breakdown was largely independent of the primary emission. Though its mechanism is incompletely understood its occurrence could be minimized by increasing the thermal and electrical properties of the coating.

Eisenstein (1946) investigated by x-ray diffraction methods the variations of surface and bulk composition of BaO–SrO cathodes during life under pulsed, d.c. and quiescent (no current drawn) conditions. On heating at 800–875° c. for 1200 hours without drawing current a surface layer largely of pure SrO was developed while the bulk composition was still about 60 mol % SrO (cf. §2.2). The surface loss of barium occurring solely by temperature treatment was found to be independent of the purity of the nickel base, but the bulk loss was strongly dependent on the presence of trace impurities in the base indicating that the barium loss must be due partly to the removal of free barium formed by the reducing action of core impurities and not wholly to evaporation of barium oxide. The effect on the bulk composition of drawing direct current for 40 hours at about 900° c. was negligible compared with that of the temperature treatment with *pure* nickel bases. However, all cathodes, whether on pure, commercial or chromium-containing nickel bases, which had been run to the 'end of life' both under d.c. and pulsed conditions showed an overall SrO content around $95 \pm 4\%$ indicating almost complete loss of barium.

Widell (1946) found the optimum composition for pulsed emission to be 50:50 mol % SrO–BaO as with normal emission (§3.3), but Eisenstein (1946 b) reports that after high current density ageing pulse emission values around 60 amp/cm² can be obtained for pure BaO cathodes—approximately twice those with very pure SrO similarly treated and equivalent to those for BaO–SrO cathodes after normal ageing. Eisenstein therefore concludes that the effect of composition on the pulse emission during life remains problematical.

Finemann and Eisenstein (1946) using x-ray diffraction methods established the formation of definite interfacial compounds with oxide cathodes treated under pulsed conditions (cf. §3.32). With pure nickel cores they obtained faint but unidentifiable interfacial patterns. With 2–5% silicon in the nickel, barium silicate was identifiable in the interface with coatings of BaO or with BaO–SrO and strontium-silicate with coatings of SrO, while with nickel containing 5% chromium or chromium-plated nickel evidence for the formation of new chromium compounds was obtained. (In a later brief communication, Eisenstein (1947) stated that *ortho*-silicates were found, in agreement with Rooksby's observations

(1947) for normal cathodes.) These authors also carried out probe measurements on BaO–SrO coatings on pure and chromium-plated nickel bases operating at temperatures of 750–850° c.; with microsecond pulses the interfacial resistances (R_i) and the bulk resistances (R_c) for a coating thickness equivalent to 10 mg/cm^2) were determined. It was found that R_i was in general considerably greater than R_c, and also showed a stronger temperature dependence which varied more widely for individual systems.*

Figure 18 shows the type of resistance current curves obtained by Finemann and Eisenstein. Coomes (1946) reports that similar results were also obtained by Buck by calculation from the observed pulse temperature rise and the power input. Finemann and Eisenstein drew attention to the resemblance of the shapes of these curves and also of their temperature dependence to those obtained for the blocking-layer type of oxide rectifier (Mott, 1939). To explain the mechanism of pulse emission they revived the conception of Reimann and Murgoci (1930) that a high core-coating interfacial resistance is a limiting condition of normal emission. Coomes pictures the layer structure for the oxide coated cathode as shown in figure 19. To explain the high emission current at microsecond pulses he postulates that the conduction band of the semiconductor (filled from the impurity levels) serves as a reservoir for the withdrawal of large transient currents at the surface; the very high fields produced in the 'blocking' layer can overcome its 'blocking' effect, and, either by the Schottky effect or field emission, cause copious emission from the base metal. The selection and pre-treatment of the base metal, rate, type and extent of interfacial compound formation (affected by base metal impurities), rather than the coating composition itself, then determine the pulsed e mission. Though evidence for a barium layer is inconclusive he considers that the surface conditions can also limit the pulsed emission under certain conditions.

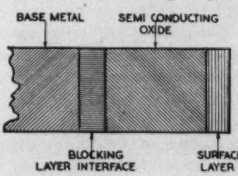


Figure 19. Layer model for oxide cathodes (after Coomes, 1946).

Wright (1947) adduces evidence from his probe measurements under D.C. conditions to show that both the phenomenon of decay observed by Sproull and that of sparking with pulsed cathodes may be direct consequences of the presence of a barrier layer at the core coating interface (see footnote on p. 319). In contrast to Sproull, who postulates that the migration of Ba^{++} ions to the core leads to a depletion of barium at the surface to which the decay in emission is due, Wright considers the more important factor is a change in the nature of the barrier layer accompanying this migration. On his evidence that the voltage drop across the cathode when drawing space-charge-limited current is some tens of volts, Wright considers that the sparking phenomenon is explicable as dielectric

* Eisenstein in a summarized communication (1947 b) reported that the observed deviation from the Langmuir-Child straight line at high current densities with microsecond pulses occurs because the potential drop across the interface and coating is not negligible compared with the applied anode potential. According to Mutter (1947) the actual potential drop determined by probe measurements accounts for the observed deviation at >800° c. Loosjes and Vink (1947) with 10^{-4} sec. pulses and current densities 5–10 amp/cm^2 have also found high potential drops across the coating—from 50–200 v. at 900–1000° K.

breakdown of the barrier layer of width 10^{-5}–10^{-4} cm. across which the bulk of the potential drop occurs. Variations in the barrier width from cathode to cathode can then lead to the observed variations in breakdown current (Coomes's sparking point) and voltage, and the observed decrease with increasing temperature of the breakdown voltage is explicable if the dielectric strength decreases with increasing temperature.

5.3. *Secondary electron emission under electron bombardment*

The maximum secondary electron emission coefficient * δ_m, defined as the maximum value on the δ/voltage curve, where δ is the number of secondaries produced per incident primary electron depends on the degree of activation of the oxide cathode. For well activated BaO–SrO cathodes at room temperature, values of 6·0 (Bruining and de Boer, 1939), 2–3 (Morgulis and Nagorsky, 1938), 4–7 (Pomerantz, 1946) and 3–6 (Megaw, 1946) have been reported. Much higher values are, however, obtained at higher temperatures.

Morgulis and Nagorsky observed a five-fold increase in δ_m at 850° K., the increase in secondary current at a temperature T° K. above that at room temperature (ΔI_{sec}) being represented over the temperature range 300–850° K. by

$$\Delta I_{sec} = A \exp\left[-E/kT\right], \qquad \ldots\ldots(36)$$

where A is a constant, and E corresponds to an energy ~0·35 ev. Similar results were obtained by Timovev and Aranovitch (1940) using thin evaporated and oxidized barium layers. Morgulis (1941, §4, p. 329) attempted a comparison of the photoelectric, thermionic and secondary emission properties of various simple oxide systems.

Pomerantz (1946 a, 1946 b) investigated the variation of δ_m with temperature in conjunction with pulsed emission measurements using magnetron-type cathodes (BaO–SrO oxides on nickel bases, pulsed at ~10 amp/cm^2 as part of the ageing treatment). The variation with temperature in the 'anomalous' yield $\Delta\delta$ ($\Delta\delta$ is the difference between δ_m at temperature T and δ_m at room temperature) was given by

$$\Delta\delta_m = A' \exp\left[-E'/2kT\right], \qquad \ldots\ldots(37)$$

where E' ranged from 0·9–1·5 ev. depending on the degree of activation. The extrapolated yields at 850° c. could be considerably in excess of 100. Pomerantz found no quantitative correlation between secondary and pulsed thermionic emission, but life tests showed there was a qualitative relation, the high temperature δ values being particularly susceptible to changes. High temperature flashing caused a sharp decrease both in δ_m and pulsed emission, but δ_m rose appreciably even after the short pulse treatment necessary for measurement of the I/V characteristic, indicating a re-activation of the surface (cf. §2.1). δ_m values of approximately 50 were obtainable from thin barely visible films. The importance of these high δ values in the performance of magnetron cathodes is stressed by Boot and Randall (1946) and Megaw (1946); however, while Megaw states that the primary and secondary emission effects can be separated, no indications of their possible interdependence are given.

* For a discussion of secondary emission phenomena and important factors therein, reference should be made to Bruining, *Die Sekundär-Electronen-Emission der festen Körper*, Berlin, 1942.

Pomerantz was able to remove the apparent conflict with the principles of the conservation of energy raised by the high yields at elevated temperatures, by showing that the maximum in the energy distribution curve of the secondary electrons was shifted towards lower energies with increasing temperature, with the result that the *total* emergent energy tended rather to decrease with temperature.

Johnson (1944) reported the occurrence of an 'enhanced' thermionic emission from oxide cathodes after primary electron bombardment under pulsed conditions which manifested itself by a slow decrease or 'tail' in the secondary electron current persisting after termination of the bombardment. Pomerantz, however, showed that the pulsed secondary emission in the microsecond range conformed to that under steady D.C. conditions at lower temperatures, so that time effects were absent; above a certain critical temperature a 'tail' in the secondary emission current was actually observed, but this was shown to be due to space-charge-limitations within the tube, and Johnson's results were thus explained.

In the absence of time effects with microsecond pulses and from the relatively small magnitude of the potential drop (estimated by resistance current), Pomerantz concluded that the large increase in δ with temperature could not be ascribed simply to a decrease in the potential drop across the coating. Morgulis and Nagorsky and later Pomerantz drew attention to the similar temperature dependence of the conductivity of impurity semiconductors (§3.23, equation (9)). It may be noted that Morgulis and Nagorsky's E value (0·35 ev.) is of the same order as that (0·3 ev.) estimated for the inner work function by Clausing and Ludwig (1933) whereas the corresponding values of Pomerantz ($E'/2 = 0{\cdot}45$–$0{\cdot}75$ ev.) are much higher. Pomerantz pointed out that while the increase in δ with temperature might be explained by an increase in the number of conduction electrons n_1 excited from the impurity levels (his E' being then identifiable with ΔE in figure 6), n_1 was small compared with the number of bound electrons even at high temperatures. Moreover, Fröhlich (1936) demonstrated the high improbability that a secondary electron emerging in the same direction as the incident primary could originate from the free conduction band. On these grounds Pomerantz therefore concluded that at least at low temperatures the main source of the secondary electrons must be the bound electrons, and that a new model of the process of secondary electron emission from such an impurity semiconductor was required.

§6. CONCLUSIONS

The considerations presented in this review show that the fundamental difficulty in our understanding of the oxide cathode is due to our lack of knowledge of the way in which the excess alkaline earth metal is built into the oxide. In regarding the emission as the passage of conduction electrons from the metal, through the semiconductor, into the vacuum, Reimann and other representatives of the 'Wilson school' disregard the fact that Wilson's semiconductor theory depends on the fact that the 'impurity' forms part of the oxide lattice. Either the barium atoms (or other alkaline earth metal) would have o be present in an interstitial position, or some oxygen ions would have to be missing from their normal lattice points. The latter type of lattice irregularity, known as the

'Schottky defect', seems unlikely to occur to an appreciable extent in a substance of such small ionic conductivity as barium oxide. There is no evidence of the formation of *F*-centres. On the other hand, the relative radii of barium and oxygen ions, which are of the same order of magnitude, make it appear equally unlikely that the barium should take up interstitial positions in the oxide lattice, as has been suggested for zinc oxide, for example (Mott and Gurney, 1940, p. 166, 'Frenkel defect'). From purely theoretical considerations it thus appears that the application of the Wilson model to BaO is not built on a very secure foundation. The de Boer concept, which regards the excess metal as adsorbed on 'internal surfaces' as well as the outer surface of the oxide, has more to recommend itself. de Boer's definition of the 'Lockerstellen' at which the impurity atoms are supposed to occur, is sufficiently vague to allow almost any interpretation to meet any particular case.

At first sight, the evidence put forward by Heinze and Wagener from electron optical and contact potential measurements, and by Nisibori and Kawamura, and Reimann and Treloar, from conductivity data, appears to be in contradiction to de Boer's concept of the emission as a process of thermal ionization of individual adsorbed atoms. The experimental results show that changes in activity of the oxide cathode are in general not due to any change in the surface condition, but primarily to a change in the inner work function of the oxide. On de Boer's theory, we should expect the A-value to increase during activation, whereas ϕ_e should remain constant, which, as Heinze and Wagener point out, is contrary to the immediate observations. However, Heinze and Wagener fail to recognize that in de Boer's theory the constant A depends not only on the covering of the surface with barium, but also on the supplementing of electrons through the coating. It has been suggested by de Boer himself, that an increase in the impurity concentration should cause a decrease in the ionization energy of the internally adsorbed atoms. The conductivity will increase, and hence also the constant A_i. Thus, the experimental observations can be satisfactorily interpreted on the basis of de Boer's theory, if we assume that changes in activity are accompanied by a change in the concentration of internally adsorbed Ba-atoms, but not, in the first instance, of the number of Ba atoms adsorbed on the outer surface. In the revised German edition of his book, de Boer has himself drawn attention to this flaw in the arguments of Heinze and Wagener.

There is no reason to doubt the applicability of the potential energy curves on which de Boer's theory depends, but a more rigorous derivation and an experimental determination of the energy difference involved would put the theory on a firmer foundation.

It is difficult to see how the 'patch-theory' interpretation of the non-saturation phenomenon can be reconciled with de Boer's concept of individual Ba atoms adsorbed on the surface. It may be that the effective radius of an adsorbed atom is considerably larger than the actual atomic radius, so that there may still be patches of varying work function of the right order of size for the patch theory to apply. If this were so, Sproull's theory of the short time decay and the accompanying decrease of the abnormal non-saturation effect would fit in better than appears at first sight with Veenemans' concept of the 'life failure' and with the general theory.

The use of the theoretical A_0 value, 120 amp. cm^{-2} deg^{-2}, in determining ϕ-values for oxide cathodes as adopted by Wagener's school seems of doubtful value. It appears wiser in view of the uncertainties involved to retain a variable A-value incorporating any variations in true emitting areas, linear temperature dependence of the work function, and so on.

The more recent war-time researches on pulsed emission have revived Reimann's original conception of a potential barrier at the core-coating interface and produced almost irrefutable evidence for the existence and possible influence of interfacial compounds. It is too early yet to attempt to draw on these results and those on the high temperature secondary emission properties for an extension of the general scheme of electron emission from oxide cathodes.

ACKNOWLEDGMENT

We wish to thank Mr. J. A. M. van Moll and the directors of Philips Electrical Ltd. (formerly Philips Lamps Ltd.) for permission to publish this review.

APPENDIX

List of recurrent symbols used

(excluding normal standard symbols)

A = General constant in Richardson equation.

A_0 = Theoretical value of A for metals = 120 amp. cm^{-2} deg^{-2}.

E_0 = Critical energy above the zero of the Fermi-surface in the core metal (or oxide).

E_1 = Energy above zero of impurity levels in semiconductor.

E_2 = Energy above zero of bottom of conduction band in semiconductor.

E_∞ = Energy above zero of electron *in vacuo* at infinite distance from the surface.

ΔE = Energy difference between bottom of conduction band and impurity level in semiconductor ($=E_1-E_2$).

ΔE_1 = Energy difference between bottom of conduction band and top of filled band in semiconductor.

i_c = Conduction current through oxide coating.

i_s, i_s^0 = Saturated emission current at applied field and zero field respectively amp. cm^{-2}).

N = Number per volume V corresponding to n, number per cm^3 with suffixes as under below.

n = Total number of electrons per cm^3.

n_1 = Number of free electrons per cm^3 in conduction band of semiconductor.

n_2 = Number of emitted electrons per cm^2.

n_a = Total number of atoms or ions of one kind (ion pairs) per cm^3 of crystal lattice.

n_i = Number of impurity levels per cm^3, or number of excess metal (Ba) atoms/cm^3.

n_i' = Number of interstitial positions per cm^3 of crystal lattice.

n_s = Number of barium atoms per cm^2, surface to form complete monolayer ($\theta=1$).

ΔU = Energy difference between bottom of conduction band in oxide and surface of Fermi distribution in metal.
V = Volume or potential difference.
V_A = Applied anode potential.
δ_m = Secondary electron emission coefficient.
σ = Specific conductance.
ϕ = Total thermionic work function in ev., generally at zero field and some temperature above absolute zero.
ϕ_0 = Value of ϕ specifically at absolute zero.
ϕ_e = ϕ-value determined experimentally from the slope of the Richardson line.
ϕ_i = Inner work function in ev.
ϕ_s = Outer or surface work function in ev.

REFERENCES

ALBRICHT, W., 1931, *Physica*, **11**, 146.
ANDRIANOV, A. M., 1944, *Bull. Acad. Sci. U.R.S.S., Sér. Phys.*, **8**, 290 ; 1945, *J. Phys. U.S.S.R.*, **9**, 60.
ARS'ENEVA, A. N. and KURTSCHATOW, B. W., 1933, *J. Exp. Theor. Phys. U.S.S.R.*, **3**, 194.
BARDEEN, J., 1947, *Phys. Rev.*, **71**, 717.
BARRER, R. M., 1941, *Diffusion in and through Solids* (Cambridge University Press).
v. BAUMBACH, H. H. and WAGNER, C., 1933, *Z. phys. Chem.*, B, **22**, 199 ; 1934, *Ibid.*, B, **24**, 59.
BECKER, J. A., 1929, *Phys. Rev.*, **34**, 1323 ; 1931, *Trans. Amer. Electrochem. Soc.*, **59**, 207 ; 1935, *Rev. Mod. Phys.*, **7**, 95.
BECKER, J. A. and MUELLER, D. W., 1928, *Phys. Rev.*, **31**, 431.
BECKER, J. A. and SEARS, R. W., 1931, *Phys. Rev.*, **38**, 2193.
BENJAMIN, M., 1935, *Phil. Mag.*, **20**, 1.
BENJAMIN, M. and ROOKSBY, H. P., 1933, *Phil. Mag.*, **15**, 810.
BERDENNIKOWA, T. P., 1932, *Phys. Z. Sowjet*, **2**, 77.
BLEWETT, J. P., 1939 a, *J. Appl. Phys.*, **10**, 668 and 831 ; 1939 b, *Phys. Rev.*, **55**, 713.
DE BOER, J. H., 1935 a, *Electron Emission and Adsorption Phenomena* (Cambridge University Press) ; 1935 b, *Chem. Weekblad*, **32**, 106 ; 1937, *Elektronenemission und Adsorptions erscheimungen* (Leipzig).
DE BOER, J. H. and VAN GEEL, W. C. H., 1935, *Physica*, **2**, 286.
BOOT, A. H. and RANDALL, J. T., 1946, *J. Instn. Elect. Engrs.*, **93**, IIIA, 928.
BORZYAK, P., 1939, *J. tech. Phys. U.S.S.R.*, **9**, 2032.
BRATTAIN, W. H., 1933, *Phys. Rev.*, **43**, 745.
BRONSTEIN, M., 1932, *Phys. Z. Sowjet*, **2**, 28.
BRUINING, H. and DE BOER, J. H., 1939, *Physica*, **6**. 823.
BURGERS, W., 1933, *Z. Phys.*, **80**, 352.
CLAASENS, A. and VEENEMANS, C. F., 1933, *Z. Phys.*, **80**, 342.
CLAUSING, P. and LUDWIG, R., 1933, *Physica*, **13**, 193.
COMPTON, K. T. and LANGMUIR, I., 1930, *Rev. Mod. Phys.*, **2**, 123.
COOMES, E. A., 1946, *J. Appl. Phys.*, **17**, 647.
DARBYSHIRE, J. A., 1938, *Proc. Roy. Soc.*, A, **50**, 635.
DAVISSON, C. and GERMER, L. H., 1924, *Phys. Rev.*, **24**, 666.
DAVYDOV, B., 1939, *J. Exp. Theor. Phys. U.S.S.R.*, **9**, 451.
DETELS, F., 1927, *Jb. drahtl. Telegr.*, **30**, 10 and 52.
DÜNWALD, H. and WAGNER, C., 1933, *Z. phys. Chem.*, B, **22**, 212.
DUSHMAN, S., 1930, *Rev. Mod. Phys.*, **2**, 381.
EISENSTEIN, A., 1946 a, *J. Appl. Phys.*, **17**, 439 ; 1946 b, *Ibid.*, 654 ; 1947 a, *Phys. Rev.*, **71**, 473 ; 1947 b, *Ibid.*, **72**, 531.
ESPE, W., 1927 a, *Wiss. Veröff. a. d. Siemens Konzern*, **5**, 29 ; 1927 b, *Ibid.*, **5** 46.
FAN, Y. H., 1942, *Phys. Rev.*, **62**, 388 ; 1943, *J. Appl. Phys.*, **14**, 552.

FELDMAN, W., 1943, *Phys. Rev.*, **64**, 113.
FINEMAN, A. and EISENSTEIN, A., 1946, *J. Appl. Phys.*, **17**, 663.
FOWLER, R. H., 1933, *Phys. Z. Sowjet*, **3**, 507 ; 1936, *Statistical Mechanics* (Cambridge University Press).
FOWLER, R. H. and WILSON, A. H., 1932, *Proc. Roy. Soc.* A, **137**, 503.
FRANCK, J., 1930, *Z. Elektrochem.*, **36**, 581.
FRITSCH, O., 1935, *Ann. Phys., Lpz.*, **22**, 375.
FRITZ, H., 1935, *Mikrochemie*, **17**, 191.
FRÖHLICH, H., 1936, *Elektronen Theorie der Metalle* (Berlin).
GÄRTNER, H., 1935, *Phil. Mag.*, **19**, 82.
GEHRTS, A., 1930, *Z. tech. Phys.*, **11**, 246.
GILL, E. W. B., 1937, *Phil. Mag.*, **24**, 1093.
GISOLF, J. H., 1942, *Nederlandsch Tijds. Natuurkunde*, **9**, 49.
GUDDEN, B., 1931, *Ber. Phys. Soc. Erlangen*, **62**, 28.
GUNDERMANN, J., HAUFFE, K. and WAGNER, C., 1937, *Z. phys. Chem.*, B, **37**, 148.
GUNDERMANN, J. and WAGNER, C., 1937. *Z. phys. Chem.*, B, **37**, 157.
HAMAKER, H. C., BRUINING, H. and ATEN, A. H. W., Jr., 1947, *Philips Res. Rep.*, **2**, 171.
HARTMANN, W., 1936, *Z. Phys.*, **102**, 702.
HEINZE, W., 1933, *Ann. Phys., Lpz.*, **16**, 41.
HEINZE, W. and WAGENER, S., 1936, *Z. tech. Phys.*, **17**, 645 ; 1938, *Z. Phys.*, **116**, 164.
HERMANN, G., 1938, *Diss. T. H. Berlin* (Hermann and Wagener, 1944, p. 179).
HERMANN, G. and WAGENER, S., 1943, *Die Oxydekathode* I ; 1944, *Die Oxydekathode* II (J. A. Barth : Leipzig).
v. HEVESY, G., 1928, *Handb. d. Physik.*, **13**, 263.
v. HEVESY, G. and SEITH, W., 1928, *Z. Phys.*, **56**, 790.
HODGSON, B., HARLEY, L. S. and PRATT, O. S., 1929, *J. Instn. Elect. Engrs.*, **67**, 762.
HORTON, F., 1906, *Phil. Mag.*, **11**, 505.
HUBER, H., 1941, *Diss. Univ. Berlin* (Hermann and Wagener, 1944).
HUBER, H. and WAGENER, S., 1942, *Z. tech. Phys.*, **23**, 1.
HULL, A. W., 1939, *Phys. Rev.*, **56**, 86.
HUXFORD, W. S., 1931, *Phys. Rev.*, **38**, 379.
ISENSEE, H., 1937, *Z. phys. Chem.*, B, **35**, 309.
JACOBS, H., 1946, *J. Appl. Phys.*, **17**, 596.
JOFFE, A. V., 1933, *Act. Scient. Ind.*, **87**, 22 ; 1946, *J. Phys. U.S.S.R.*, **10**, 49.
JOFFE, A. V. and JOFFE, A. F., 1940, *J. Phys. U.S.S.R.*, **2**, 283 ; 1941, *Bull. Acad. Sci., U.R.S.S., Sér. Phys.*, **5**, 550.
JOFFE, J., 1945, *Elect. Commun.*, **22**, 217.
JOHNSON, J. B., 1944, *Phys. Rev.*, **60**, 352.
JUSÉ, W. P. and KURTSCHATOW, B. W., 1932, *J. Exp. Theor. Phys. U.S.S.R.*, **2**, 309 ; 1933, *Phys. Z. Sowjet*, **2**, 453.
KNIEPKAMP, H. and NEBEL, C., 1932, *Wiss. Veröff. a. d. Siemens Konzern*, **11**, 75.
KNIGHT and HERBERT, 1946, *J. Instn. Elect. Engrs.*, **93**, IIIA, 949.
KOLLER, L. R., 1925, *Phys. Rev.*, **22**, 671.
KROCZEK, J. and LÜBKE, E., 1930, *Wiss. Veröff. a. d. Siemens Konzern*, **9**, 252.
LANGMUIR, I., 1920, *Gen. Elect. Rev.*, **23**, 503 and 589.
LE BLANC, M., SACHSE, H. and SCHRÖPEL, H., 1933, *Ann. Phys., Lpz.*, **11**, 334.
LIEBOLD, W., 1941, *Diss. Univ. Berlin* (Hermann and Wagener, 1944, p. 228).
LINFORD, L. B., 1930, *Phys. Rev.*, **30**, 1100 ; 1931, *Ibid.*, **37**, 1018 ; 1933, *Rev. Mod. Phys.*, **5**, 34.
LOOSJES, R. and VINK, H. J., 1947, *Philips Res. Rep.*, **2**, 190.
LOWRY, E. F., 1930, *Phys. Rev.*, **35**, 1367.
LYASHENKO, V. I., 1942, *J. Exp. Theor. Phys. U.S.S.R.*, **12**, 281.
MADDOCK, A. J., 1935, *Phil. Mag.*, **19**, 422.
MAURER, R. J., 1945, *J. Appl. Phys.*, **16**, 563.
MEGAW, E. C. S., 1946, *J. Instn. Elect. Engrs.*, **93**, IIIA, 977.
MEYER, W. and NELDEL, H., 1937, *Phys. Z.*, **38**, 1014.
MEYER, W and SCHMIDT, A., 1932, *Z. tech. Phys.*, **13**, 137.
MILLIS, W. T. and LOWRY, E. T., 1934, *Phys. Rev.*, **45**, 746.

MÖGLICH, F., 1938, *Z. Phys.*, **109**, 503.
MÖNCH, G., 1933, *Z. Phys.*, **83**, 247.
MOORE, G. E. and ALLISON, H. W., 1941, *J. Appl. Phys.*, **12**, 431.
MORGULIS, N., 1941, *Bull. Acad. Sci. U.R.S.S., Sér. Phys.*, **5**, 536.
MORGULIS, N. and NAGORSKY, A., 1938, *J. tech. Phys., U.S.S.R.*, **8**, 848.
MOTT, N. F., 1938, *Proc. Camb. Phil. Soc.*, **34**, 568 ; 1939, *Proc. Roy. Soc.*, A, **171**, 27.
MOTT, N. F. and GURNEY, R. W., 1940, *Electronic Processes in Ionic Crystals* (Oxford University Press).
MUTTER, W. E., 1947, *Phys. Rev.*, **72**, 531.
NISIBORI, E. and KAWAMURA, H., 1940, *Proc. Phys.-Math. Soc., Japan*, **22**, 376.
PEKAR, S. I., 1940, *J. Exp. Theor. Phys. U.S.S.R.*, **10**, 1210 ; 1941, *Bull. Acad. Sci. U.R.S.S., Sér. Phys.*, **5**, 422.
PHIPPS, T. E. and LESLIE, R. T., 1928, *J. Amer. Chem. Soc.*, **50**, 2412.
PODSZUS, E., 1933, *Z. Elektrochem.*, **39**, 75.
POMERANTZ, M. A., 1946 a, *J. Franklin Inst.*, **241**, 415 and **242**, 41 ; 1946 b, *Phys. Rev.*, **70**, 33.
PRESCOTT, C. H. and MORRISON, J., 1938, *J. Amer. Chem. Soc.*, **60**, 3047 ; 1939, *Rev. Sci. Instrum.*, **10**, 36.
REIMANN, A. L., 1934, *Thermionic Emission* (London : Chapman and Hall).
REIMANN, A. L. and MURGOCI, R., 1930, *Phil. Mag.*, **9**, 440.
REIMANN, A. L. and TRELOAR, L. R. G., 1931, *Phil. Mag.*, **12**, 1073.
REYNOLDS, N. B., 1930, *Phys. Rev.*, **35**, 158.
RICHARDSON, O. W., 1914, *Phil. Mag.*, **28**, 633.
RICHARDSON, O. W. and v. LAUE, M., 1918, *Jb. Radioakt.*, **15**, 205, 207.
RICHARDSON, O. W. and YOUNG, A. F., 1925, *Proc. Roy. Soc.*, A, **107**, 377.
ROOKSBY, W. P., 1947, *Nature, Lond.*, **159**, 609.
ROOKSBY, W. P. and STEWARD, E. J., 1946, *Nature, Lond.*, **157**, 548.
ROSE, A., 1936, *Phys. Rev.*, **49**, 838.
ROTHE, H., 1926, *Z. Phys.*, **36**, 737.
ROUSSEAU, E., 1934, *Chimie et Industrie*, Spec. no. 775 (April).
SCHADE, O., 1943, *Proc. Inst. Radio Engrs.*, **31**, 341.
SCHOTTKY, W., 1914, *Phys. Z.*, **15**, 872 ; 1935, *Naturwissenschaften*, **23**, 115.
SCHOTTKY, W. and HARTMANN, 1933, *Z. tech. Phys.*, **16**, 312.
SCHOTTKY, W. and SPENKE, E., 1939, *Wiss. Veröff. a. d. Siemens Werke*, **18** (3), 1.
SCHRIEL, M., 1937, *Z. anorg. Chem.*, **231** (4), 313.
SCHULMAN, A. R., 1940, *J. tech. Phys., U.S.S.R.*, **10**, 1173.
SEITZ, F., 1945, *J. Appl. Phys.*, **16**, 553.
SIMON, H., 1927, *Z. tech. Phys.*, **11**, 434 ; 1928, *Handb. Experimentalphysik*, **13**, (2) 305.
SMEKAL, A., 1928, *Z. Elektrochem.*, **34**, 472.
SPANNER, H. J., 1924, *Ann. Phys., Lpz.*, **75**, 609.
SPROULL, R. L., 1945, *Phys. Rev.*, **67**, 166.
TIMOVEV, A. W. and ARANOVITCH, R. M., 1940, *J. tech. Phys., U.S.S.R.*, **10**, 32.
UEHARA, Y., 1933, *J. Chem. Soc., Japan*, **60**, 133.
UEHARA, Y. and TAKAHARI, M., 1940, *Bull. Chem. Soc., Japan*, **15**, 15.
VEENEMANS, C. F., 1943, *Nederlandsch Tijds. Natuurkunde*, **10**, 1.
VILLARS, D. S. and DUSHMAN, S., 1930, *Rev. Mod. Phys.*, **2**, 433.
VOGT, W., 1931, *Ann. Phys., Lpz.*, **7**, 183.
WEHNELT, A., 1903, *Verh. dtsch. phys. Ges.*, **5**, 255.
WHEATCROFT, E. L. E., 1940, *Phil. Mag.*, **29**, 16.
WIDELL, E. G., 1946, *Bull. Amer. Phys. Soc.*, **21**, E2.
WILLIAMS, N. H. and HUXFORD, W. S., 1930, *Phys. Rev.*, **36**, 463.
WILLSHAW, W. E., RUSHFORTH, L., STAINSBY, A. G., LATHAM, R., BALLS, A. W. and KING, A. N., 1946, *J. Instn. Elect. Engrs.*, **93**, IIIA, 985.
WILSON, A. H., 1931 a, *Proc. Roy. Soc.*, A, **133**, 485 ; 1931 b, *Ibid.*, **134**, 277 ; 1939, *Semiconductors and Metals* (Cambridge University Press).
WRIGHT, D. A., 1947, *Proc. Roy. Soc.*, A, **190**, 394.

PHYSICS IN GLASS TECHNOLOGY

By H. MOORE,
Department of Glass Technology, University of Sheffield

§1. INTRODUCTORY

THE application of physics on any wide scale to technical and scientific problems concerning glass and glass manufacture dates back little more than a quarter of a century. Before 1914 one might find here or there some simple form of strain-indicating device in use as a check on annealing, or occasionally, some form of pyrometric equipment for measuring furnace or kiln temperatures; in most sections of the glass industry, however, annealing was judged to be satisfactory if the manufactured article did not "fly" during cooling, and kiln temperatures were estimated by eye.

Control of composition by chemical analysis of raw materials and of finished glass was already well established by 1914 in certain sections of the industry, though it was by no means a general practice. Some study had been made, also, of the ways in which certain physical properties of glasses could be altered by changes of composition. Of these properties, density, refractive index and expansion coefficient were the ones most extensively studied, and already some attempts had been made to develop generalized relations between them and the proportions in which various constituents were present in the glasses. This was virtually the limit to which physics had been applied to glass and the results consisted, in essence, of tables giving certain physical properties of glasses, some of which could be regarded as forming, or belonging to, fairly definite groups or series, with a few formulae, not entirely reliable, which indicated how these physical properties could be calculated from the percentage composition. Most of this systematic observation of physical properties had been done at Jena, in parallel with determinations of chemical durability carried out in connection with the various new glasses developed by Schott for optical and chemical purposes. Very little had been done elsewhere, either in this country or America, and, for the most part, each firm limited itself to the production of only one or two types of glass, the constancy or variability of which depended on the skill of workmen who simply "knew how". It should be said, at once, that the comparatively small variations over long periods was ample testimony to that skill and knowledge.

Apart from plate-glass manufacture, most processes in the industry had been almost entirely manual or manually operated, mechanization having been limited virtually to the introduction of simple presses and blowing machines, usually hand-fed. In these processes, whether entirely manual or semi-mechanized, considerable variations in the properties of the glass could be allowed for by the skilled blower or machine-operator. If the glass was too "hard" to be worked at the usual working temperature it was not allowed to cool quite so much as usual before blowing or pressing after gathering or ladling. In the post-1918 period,

however, methods developed more and more towards fully automatic production, and the machines were obviously incapable of applying the compensations and adjustments which had been made by the skilled worker. As a result, closer control became necessary on composition and temperature and on every other factor which might influence the homogeneity and properties of the molten glass. It became necessary, also, to devise means for the continuous and automatic annealing of the finished glass, whatever its form, in order to ensure the smooth flow of the greatly increased output from the automatic machines; attention had to be directed, therefore, to the design of annealing lehrs and to the control of lehr temperatures to ensure adequate annealing in the minimum possible time.

The introduction of temperature indicating and recording instruments was probably the first wide-spread application of physics into glass technology. It was soon found, however, that the measurement of furnace and lehr temperatures was not an entirely simple matter. A thermocouple put into the crown of a furnace did not give the same temperature reading as a disappearing-filament pyrometer focused on the furnace walls, while a radiation pyrometer directed on to the surface of the molten glass gave a different reading again. In a lehr there appeared to be no satisfactory method of measuring the temperature of the articles in process of annealing. Merely to put a fixed thermocouple into the lehr somewhere near the moving glass was very soon seen to be useless, and it was quite impracticable to put a travelling thermocouple into actual contact with the glass. Furthermore, little was known about the temperature-time schedule required for good annealing.

Owing to these difficulties and to the inconsistencies between the readings of different types of pyrometers, all attempts to measure temperatures were abandoned by many firms, and pyrometers were frequently described as "unreliable". Among the more enlightened firms in the industry, however, it was recognized that some form of pyrometer was essential if the eye of the skilled furnaceman was ever to be replaced by an index and scale, and they persevered in their efforts to make the pyrometer readings "mean something" in relation to glass melting, processing and annealing. Not infrequently this was done by putting a pyrometer in some convenient position and determining, by trial, the temperature it recorded when working conditions were correct, thereafter using its indications as a check on the constancy with which those conditions were maintained, irrespective of the precise relation between that temperature and the true temperature of the glass. Investigations were begun, also, to determine with reasonable precision the temperature at which melting would be carried out most effectively; the rate of release of strain at various temperatures was also investigated with a view to developing the best temperature-time schedule for rapid but adequate annealing, and so forth.

It would, however, be laborious to attempt to trace in chronological order the progress of physics in glass technology. Out of the apparently simple problem of temperature measurement there grew a realization of a need for much further knowledge concerning the various manufacturing processes, and concerning the physical properties of glass on which the processes themselves depend. As was inevitable, with new knowledge new "vistas of ignorance" were

discovered, requiring more and more physical knowledge and experiment for their proper exploration, and the physicist eventually was given a place in the industry beside the chemist, the engineer and the skilled glass-maker.

The complexity of the problems of glass technology makes it necessary for many of them to be considered from a number of angles, and it is impossible for any informative report of the progress of physics in glass to omit some discussion of the chemical and engineering aspects. Possibly the best mode of approach is to describe in outline the processes by which glass is made and shaped, then to go on to deal with the ways in which physics and physical investigations have been applied to the elucidation of the problems arising, with such reference as may be necessary to their chemical and engineering aspects.

§2. OUTLINES OF GLASS MANUFACTURE

Glasses are made by melting together various inorganic materials which, with one or two exceptions, are broken down to acidic or basic oxides which combine together during the melting process. Silica is the main acidic constituent of all the more usual glasses; in making glass the silica is heated in the presence of fluxes, the most common of which are the alkalis, soda and potash, with lime or a mixture of lime and magnesia to render the resulting glass insoluble in water. Barium oxide, zinc oxide and sometimes strontium oxide are also used to some extent, replacing some or all of the lime and magnesia more commonly used. Lead oxide, also, is an important fluxing agent, but it seems to behave somewhat differently from the alkalis and alkaline earths, and is capable of being used in very large proportions, up to 90 % of the total glass if desired. Acidic fluxing agents, used in some cases to assist melting or to impart special qualities to the glass, are boric oxide, fluorides or silico-fluorides, and phosphoric oxide, which is added as phosphoric acid or alkali phosphate. Alumina is another common constituent of glasses, though not usually in amounts exceeding 2 % to 3 %; it may be regarded broadly as a neutral constituent, or perhaps it would be best described as amphoteric, combining as an acid with the alkalis, soda and potash, but also capable of combining as a base with the silica and particularly with boric oxide and phosphoric acid when those are present in the glass.

The various ingredients to be melted together are put into a crucible or pot or " tank " of refractory material capable of withstanding the high temperature needed to melt the glass, and also capable of resisting satisfactorily the fluxing action of the molten fluxes and of the molten glass.

In " pot-melting " the pots are placed inside a furnace and the temperature is raised until the ingredients have melted down to form a substantially homogeneous glass. The glass is then maintained for some time at a higher temperature (1350° c. to 1600° c. according to the type of glass being made) to allow entrapped bubbles of gas to escape, after which it is allowed to cool to " working temperature ", i.e. the temperature at which its viscosity is suitable for whatever method of working may be necessary to produce the type of finished article required.

In " tank-melting " the furnace is itself used to contain the molten glass. The floor and walls are constructed of refractory blocks having a high resistance to attack by molten glass, and above glass level the walls are continued vertically

some two or three feet, usually in silica brickwork, the whole being covered with an arched " crown ", also made of silica bricks.

Modern furnaces of either type are heated either by gas or oil, air being supplied through separate ducts. Preheating of the air, either by recuperators or regenerators, is universal, and regenerator heating of the ingoing gas is used in all furnaces supplied by producer gas, though not usually when " town's gas " or " natural gas " is used as the fuel.

The molten glass may be shaped by blowing, pressing, drawing or rolling. Many articles are " hand-blown " either " free " or into moulds; " machine-blowing " into moulds is used extensively for mass production of bottles, jars and many other forms of containers. Pressing is obviously a moulding operation in essence, though the moulding machines may be hand operated. Drawing may be entirely manual or entirely mechanical; it is used for the production of glass-tubing and rod, and also for making modern " flat-drawn " sheet glass. Rolling is obviously a mechanical process, usually fully mechanized.

After shaping, the glass must be annealed in order that it may be reasonably free from strain when cold. Modern production requires that the annealing should be carried out as rapidly as possible, and the usual practice is to convey the glass mechanically through an annealing furnace or " lehr " along the length of which the temperature is adjusted to suit the weight and dimensions (particularly thickness) of the glass passing through, and the speed at which it travels.

Cutting, grinding, polishing and other operations may follow, according to the type of article being produced, with, finally, warehousing and packing for transport.

Each of these processes, from the preparation of the refractory pots and the mixing of the raw materials to the warehousing and packing, presents its own problems, to the solution of which scientific methods, including physical methods, have been applied.

§3. APPLICATIONS OF PHYSICS IN GLASS MANUFACTURE

(i) *Refractory manufacture*

(a) *Glass-melting refractories.* The essentials of a refractory material for use in glass-making are :—

(1) that it should be capable of being shaped in the cold, and be strong enough to withstand subsequent handling without fracture or deformation;

(2) that it should be infusible at the highest temperature reached in the furnace during melting, should not crack or deform whilst being heated to that temperature, and should be strong enough at that temperature to withstand, without deformation or cracking, the combined load due to its own weight and to the weight of the glass it has to support;

(3) that it should be adequately resistant to attack by the fluxing action of the raw " batch " materials and of the molten glass.

The majority of refractory materials used for glass melting consist of clays or clay mixtures (i.e. of various types and mixtures of aluminium silicates) sometimes containing free silica. The pure aluminium silicates are, in general, highly

infusible, but do not form strongly cohesive plastic masses when " wetted up ". Fluidity, plasticity and " wet " strength largely depend on the presence of small amounts of impurities in the natural clays, particularly on the amounts of alkalis present. The presence of alkalis, however, reduces the refractoriness of the clays when burnt, i.e. alkali present in or added to a clay lowers the temperature at which the clay becomes soft enough to flow or deform under small loads or under its own weight. In assessing the suitability of a clay for glass-making, therefore, a nice balance must be held between the requirements of plasticity and refractoriness.

Most of the natural clay deposits contain appreciable proportions of iron oxide which makes the clay unsuitable for glass-making if the amount is more than 2% or 3%; the ferric oxide, in addition to having some fluxing action on the aluminium silicates, is taken into solution by the glass along with the alumina and silica as the clay is gradually corroded, and, if present in large proportion, imparts an undesirable colour to the glass.

An electrolytic method of purifying clays was developed some thirty years ago. The applicability of this method depends on the fact that when clays are mixed with small proportions of soda or other alkaline " dispersing " agents, the relatively large aggregates of clay particles break down as the fine particles adsorb the alkali ions, forming a highly fluid " slip ". Since each particle of clay is now associated with an alkali ion, the particles will move towards the cathode if the slip is fed into an electrolysing bath, and a layer of fine clay begins to accumulate on the cathode.

In the commercial application of this process the slip was fed continuously into an electrolysing bath in which the cathode was made in the form of a perforated drum and covered by a layer of filter cloth. As the clay particles were carried over they were collected on the filter cloth, but the alkali ions passed through to the cathode and were swept away in the water effluent which flowed away axially inside the drum. The cathode was rotated slowly against a scraper which removed the clay as rapidly as it was deposited on the filter cloth.

The clay thus obtained was associated with only a very small proportion of the alkali originally used to disperse it; moreover, particles in the original clay which did not adsorb the alkali ions were not carried over to the cathode, but simply settled out as a sludge at the bottom of the electrolysing bath. The product was thus of a very high degree of purity and of a correspondingly highly refractory character, but stiff and lacking in plasticity owing to its very small alkali content. By adding appropriate proportions of alkali it could, of course, be brought to any degree of plasticity desired, or it could be blended with clays from natural deposits to stiffen them and to increase their refractoriness. " Electrosmosed " clay was used to an appreciable extent on the Continent, but was not produced on a commercial scale in this country, except for a short period, because the ball clay and china clay deposits in the West and South-West of England are sufficiently pure for all ordinary purposes and can be blended together, or with other clays, to produce the various degrees of plasticity and refractoriness desired.

At one time clays for pot-making were literally " trodden " before being worked to shape, in fact, this practice was comparatively common up to 15 or

20 years ago and still persists in isolated cases. The pots were built up by hand, as they still are to a very large extent, being then put aside for several months to dry naturally in slightly heated chambers or " pot-rooms ". When demands for increased production of glass made speedier pot production necessary, mechanical methods of blending were introduced, but the actual shaping of the pots by mechanical methods presented serious difficulties, as did also the apparently simple problem of drying the finished pots.

The pressing of pots to shape was solved by the art of the refractories chemists, but it is interesting to note that the suitability of clays for the pressing process appears to be determined in part by the nature or proportions of the various crystalline forms of aluminium silicates present in the clays. If these indications are substantiated x-ray diffraction methods of examining clays should be of considerable value in assessing the suitability of clays for making pressed pots and other refractory shapes to be used in glass-making, as it would enable unsuitable clays to be rejected without the need for undertaking the lengthy and expensive preparation work required for making practical trials.

The method developed for the rapid drying of clay pots is an interesting example of the practical application of simple physical principles to the solution of a problem which had proved baffling for many years. The considerable shrinkage which occurs when wet clay is dried causes surface cracks to develop if the moisture content of the surface layers becomes substantially lower than that of the main body of the clay. The usual practice, therefore, is to make up the pots, then to store them in gently warmed pot-rooms for periods of six to seven months, by which time they are sufficiently dry to be " burned " and put into use. The normal life of a pot such as is used in the older process of plate glass manufacture may be taken as between ten and twenty " journeys ", occupying a period of one and a half to three weeks. To allow for the regular replacement of the pots, with some margin in case of a bad " run " of breakages, this requires some fifteen to twenty pots in the drying rooms for every pot actually in the furnaces, representing a high capital outlay, not only in the pots themselves, but also in providing buildings strong enough to carry them during the drying period. There has thus always been a strong incentive to develop a rapid method of drying, and this became particularly important when demands for increased production became urgent.

It is obvious that the moisture must escape from the surface of the clay and that the rate of evaporation from the surface can be increased by raising the temperature. For a given " moisture gradient ", also, rise of temperature will increase the rate of diffusion of moisture from the interior of the clay to the surface. The problem of accelerated drying thus resolves itself into raising the temperature, but controlling the evaporation from the surface in such a way as to create but not exceed the maximum safe moisture gradient in the clay. With the problem reduced to these simple physical principles it is evident that the desired result can be obtained by heating the pots uniformly to something near 100° C. in an atmosphere in which the relative humidity can be controlled at some suitable value, probably between 85 % and 95 %. In practice the desired conditions can best be established by determining the relative humidity at which surface crazing just begins, then increasing the humidity slightly, but just enough

to give a small margin of safety. The effectiveness of this method can be gauged from the fact that the drying of pots can now be completed in a period of ten to fifteen days instead of the six to seven months needed in the usual drying process.

X-ray diffraction methods have found many useful applications in investigations of the changes which occur when clays and clay mixtures are burned at various temperatures. During burning, a clay mixture develops a crystalline structure, the nature of which depends on the composition of the clay and on the temperature to which it is fired. As the temperature is raised, mullite crystals begin to predominate, any silica or alumina in excess of the proportion corresponding with the composition of mullite ($3Al_2O_3 : 2SiO_2$) being freed. If excess silica is present it combines partially or wholly with the alkalis or other fluxes to form a glassy matrix surrounding the mullite crystals, or if alumina is present in excess it forms crystals of corundum mixed with the mullite. The glassy matrix always contains silica and alumina in addition to the fluxes present in the original clay mixture, but the proportion of glassy matrix to crystalline content of the burnt clay is determined by the proportion of fluxes present.

A structure consisting of closely interlaced needles of mullite has considerable mechanical strength at high temperatures, and the mullite crystals are also highly resistant to solution by molten glass ; corundum also, is highly refractory and offers great resistance to corrosion. The glassy matrix, however, tends to soften at high temperatures and thus reduces the " hot strength " of the burnt clay. The proportion of glassy matrix should thus be kept low by keeping down the proportion of alkalis and other fluxes in the original mixture, and the matrix should be a glass having a high softening point.

The discovery that mullite was the ultimate form of crystal developed in the burning of clays has afforded a valuable guide to the composition to be aimed at in developing clay mixtures to be used as pot-clays and tank blocks for glass-making. X-ray diffraction methods have also given valuable information concerning the extent to which crystal development in clays at high temperatures may be stimulated by the presence or addition of different proportions of alkalis, lime, magnesia and other " mineralizing " agents. If, as appears probable, X-ray methods can also be used to determine the actual proportions of the various crystalline phases and the proportion of glassy phase present in burnt refractories, the information obtained in these investigations on crystal development should be considerably extended. A knowledge of the proportions of crystalline and glassy phases will also be interesting to compare with direct measurements of strength and "creep" under load at high temperatures. These determinations of strength and creep at high temperatures are themselves examples of simple applications of physical methods to the development and testing of refractories for glass-making. Measurement of the flow of clays when "wetted up" with different proportions of water, with or without added dispersing agents, simple methods of comparing the fluidity of clay slips, and many other methods of testing clays, also have a strong physical flavour, though in the main they were devised by refractories chemists in the search for essential practical information concerning the materials with which they worked. Many of these measurements could be expressed in terms of viscosity, but the methods are so well standardized and the results can be so readily translated into practical form that there would be little

direct value in carrying out the, in some cases, rather involved experimental work which would be required.

(b) *Refractories for use as insulating materials and for other purposes.* The use of insulating materials to reduce heat losses from producer-gas ducts, furnaces and regenerators is obviously a means of securing general economy of operation, and for such purposes different kinds of refractories may be needed depending on the temperatures to which the insulating materials will be exposed on the "hot face". Laboratory investigations of the thermal properties of refractories including determinations of specific heats and measurements of thermal conductivity over wide temperature ranges, are obvious applications of physics of direct practical value in comparing the insulating properties of different types of refractories, whether for routine control of properties or for developing improved types. To obtain accurate and complete information concerning either thermal conductivity or specific heat, both of which vary very considerably over the ranges of temperature of importance in practice, would require considerable ingenuity in arranging the experiments, and experimental skill of a high order. Fortunately, perhaps, it is sufficient for most purposes to have mean values of specific heat or mean values of thermal conductivity over wide ranges of temperature and the order of accuracy demanded is low.

Refractory bricks are used as the "stuffing" or chequer work in regenerators. The object of the chequer work is to absorb heat as completely as possible from the hot gases as they leave the furnace, and to release the heat again to the ingoing gas and air when the firing of the furnace is reversed. There is no agreement among specialists in furnace design concerning the relative advantages of using refractories of high or low specific heat and high or low thermal conductivity for this purpose, and there may be an opportunity here for the problem to be solved theoretically on physical grounds, though initial costs and costs of replacement are likely to be the decisive factors in the actual choice of the bricks used most extensively.

There is, however, an interesting example of the use of physical methods to trace out the course of reactions which cause a comparatively rapid degradation of the bricks in the upper courses of the chequer work when acted on by alkali vapours and sulphurous gases from the furnace. The reactions leading to this degradation occur in the solid state, but chemical analysis fails to establish their nature because it fails to distinguish between simple absorption and actual combination. The progress of the reactions can, however, be traced out clearly by determining the nature of the crystals which develop in the bricks at various stages. The initial swelling of the bricks is due to the formation of complex "additive" compounds of well-defined crystal structure, but not highly stable at moderately high temperatures (700 to 800° C.). The tendency is, therefore, for one component (sodium sulphate) to penetrate into the cooler interior of the brick, leaving the outer layers of brick badly crazed and covered with a new crystal development resembling a brown moss, which eventually becomes detached and falls away, thus impeding the flow of air and gas by partially or eventually completely blocking the gas and air channels. The sodium sulphate which has penetrated into the interior of the brick causes further swelling and cracking, and its action is reinforced as additional amounts of the alkali vapours and sulphurous gases are brought over. Breakdown of the bricks thus continues progressively as

the affected portions of the brick become exposed to higher temperatures due to the falling away of mossy growth on the surface. The course of the reactions having been established, the next step should obviously be to seek for refractories which will react less readily in these ways, since it is virtually impossible to avoid the presence of alkali vapours and sulphurous compounds in the gases leaving the furnace.

Other examples of the applications of physics and physical methods to refractory manufacture could be cited, but those already indicated may be regarded as giving a sufficiently general picture.

(ii) *Glass melting*

The development of batch compositions and the making up of batch mixtures are primarily a matter for the chemist. In the melting of the glass, however, physical principles can be applied in so many ways and to problems so diverse in character that any exhaustive description or discussion would be difficult or impossible. Methods of assessing the thermal efficiency of a complete glass melting layout, for example, involve determinations of heat losses from gas ducts, regenerators or recuperators, and from the furnace itself. The modes of heat transfer from the furnace flames to the melting glass also require considerable study, while the most obvious types of observation which may be of value include measurements of temperature, determinations of viscosity at various temperatures, variation of specific gravity with composition and temperature, surface tension measurements, and so forth.

Information on certain of these points may be of direct value in the improvements of processes, or in securing economies in production; in other cases it may be of no great importance or value, even though it may explain some detail of established practice, or, again, it may simply represent a point of additional knowledge which may assist at some time or other in solving a production problem or in developing some new process. The progress in glass melting which has resulted from the physical method of approach may be best indicated, perhaps, by instancing some of the major investigations or groups of investigations and describing some of the results obtained.

(a) *Investigation of furnace efficiency.* From a commercial point of view the efficiency of operation of a furnace may be expressed in terms of the amount of good quality glass produced per 1000 B.Th.U. supplied by whatever form of fuel is used. Of the heat supplied a portion is used in raising the temperature of the raw "batch" and in melting it, the rest being needed to maintain the furnace temperature, i.e. in supplying heat to balance the heat lost in various ways from the furnace. The heat used in heating and melting the batch can be assumed to be proportional to the amount of batch melted. The heat required to balance losses when the furnace is maintained at this working temperature is, however, a property of the furnace itself, and is substantially independent of the rate at which glass is being made. In order to assess the efficiency of the furnace, therefore, as distinct from the efficiency of operation, it is this latter quantity which must be determined in relation to the maximum rate of melting possible in the furnace.

The problem can be reduced to a simple form if we assume that the maximum rate of melting to produce glass of good quality is proportional to the "melting

area" of the furnace. This enables the relative efficiencies of furnaces of different sizes to be compared by determining the heat loss per hour per square foot of melting area. Certain correction factors are necessary if a fair comparison is to be made between furnaces melting glasses which require different melting temperatures, or if glasses of different colours are being melted; a small correction is allowed, also, in comparing furnaces differing largely in melting area. This method of assessing the efficiency of furnace design quantitatively has been developed only within the past two or three years by the Furnace Committee of the Society of Glass Technology.

Much detailed information about the heat "balance-sheet" of furnaces has been accumulated during the past fifteen years or more by laborious and difficult observations of heat input and heat losses. These have involved estimation of the total calorific value of the fuel used, measurement of sensible heat retained by the producer gas as it entered the regenerators, and measurement of gas and air temperatures at the top of the regenerators immediately before entering the furnace. On the exit side the temperature of the burnt mixture as it left the furnace and again as it left the regenerators had to be measured. From all this, coupled with the knowledge of the specific heats of the gas, air, and burnt mixture, the heat actually developed inside the furnace could be calculated. The possible errors were large, necessitating many repeated determinations, but eventually reasonably accurate results were obtained. Comparison of these results with approximate calculations of the heat absorbed in heating and melting the raw materials, based on their specific heats, showed an enormous disparity, the heat used in converting the raw materials into molten glass being barely 10% of the heat actually generated inside the furnace. This figure was fully confirmed by calculations of the heat losses through the furnace walls, crown, and bottom-blocks, for which purposes the thermal conductivities of the various types of refractories were determined and the temperatures of their outer surfaces measured. Changes of design directed towards reduction of these losses have effected appreciable improvements, but even in the most modern of tank furnaces it is doubtful if the proportion of heat used in "melting" is as high as 20% of the total heat supplied.

(b) *Efficiency of producers and regenerators*; *design of furnace "ports"*. Detailed consideration of the overall efficiency of glass melting plant using producer-gas, however, must begin at the producers themselves. Proper control of the producers to ensure a minimum of combustible material in the ash, control of steam pressure, and so forth, may be regarded as outside the scope of this Report. Determinations of calorific value of the producer gas, in combination with chemical analysis of the products of combustion, serve as a check on the efficient operation of the producers, and thereafter the gas must be conveyed to the furnace with minimum loss of heat and with minimum loss of combustible vapours. At one time it was comparatively common for the gas ducts to be in the form of brick-built tunnels below ground level. Frequently they were of considerable length, of comparatively small cross-section, and with no protection against water soaking into the brickwork. The simplest physical considerations indicate that improvements would be obtained by making the gas ducts as short as possible, of large cross-sectional area, and by running the ducts above ground level, well-lagged, with the lagging thoroughly waterproofed. Such changes must often wait many

years before they can be put into effect, on account of the disorganization of production they involve, but there have been steady developments in this direction. The advantages to be gained by such changes are not limited to the mere retention of sensible heat by the gases. If water-vapour is picked up as the hot gas passes through damp or wet ducts, not only is the gas cooled but in addition the water-vapour must be heated as it passes through the regenerators and absorbs more heat inside the furnace itself, necessitating a greater supply of gas to obtain any specific furnace temperature. By keeping the gas hot on its way to the furnace there is the minimum loss of calorific value due to condensation of the heavy tarry vapours carried over in the gas. Moreover, increase of cross-sectional area of the ducts reduces the velocity of the gases through the ducts and so reduces the amount of fine ash, usually rich in iron oxide, which is carried into the furnace as dust. Obviously there is a limit beyond which increase of cross-sectional area may be disadvantageous; for example, the reduction in the heat loss due to using wider ducts will be off-set to some extent by the longer time taken by the gas to pass through any length of duct. The optimum diameter must be determined by trial, however, since the movement of the gas is associated with some turbulence, consequently theoretical considerations permit only a very rough estimate to be made of the diameter beyond which any further increase would be undesirable.

The study of regenerator design also offers scope for physical investigation. Determinations of regenerator efficiency involve measurement of temperature of the burnt gases as they enter the regenerator from the furnace and again as they leave the regenerator and are carried to the chimney stack. When the furnace is "reversed" similar measurements must be made on the ingoing streams of gas and air entering the regenerator and then as they leave the regenerator, immediately before they become mixed and combustion begins. The dimensions of regenerator chambers should bear some definite relation to the volumes of gas and air passing through them, but the very wide divergences met with, even in furnaces of modern design, are an indication of the large amount of further work necessary.

The advantages to be gained by insulating the regenerators to avoid loss of heat, and the importance of avoiding gas and air leaks are obvious. There are, however, other matters which could usefully be investigated, such as the best pattern in which to arrange the brickwork "lattice" so as to ensure maximum heat interchange between the bricks and the gases, and the relative advantages of using bricks of low or high specific heat and high or low thermal conductivity.

Problems of recuperator design are, perhaps, simpler. Recuperators are used mainly on furnaces fired by oil or by gas of high calorific power such as natural gas or "town's" gas (coal gas) where the thermal capacity of the fuel is small compared with that of producer gas giving the same heat output. In a recuperator the ingoing air passes through refractory tubes in a chamber through which the hot waste gases from the furnace are drawn on their way to the chimney stack. Good thermal interchange requires that the refractory tubes should have a reasonably high thermal conductivity, in fact, heat resisting metal tubes have been used instead of refractory tubes, though they have not been adopted in any large measure. In practice, air leaks are the most serious causes of low recuperator efficiency, particularly air drawn in through cracks or bad jointing in the walls of the regenerator chamber. Any air entering in this way will obviously cool the hot

furnace gases by direct mixture, so that transfer of heat to the air passing through the refractory tubes is greatly reduced. Such leaks are not readily detected. Temperature measurements of the burnt gases as they enter and leave the recuperator, in addition to being difficult may actually be misleading, since a large difference of temperature due to cooling by cold air leaking in might be taken to indicate high efficiency due to more effective interchange through the walls of the refractory tubes. As a result, analysis of the gases as they leave the recuperator on their way to the chimney stack has been the general method of testing for leaks. The possibility of applying physical methods has, however, received some consideration, and there would be obvious advantages in the development of a method by which any leak could be indicated by a corresponding change in some physical quality which could be continuously recorded.

(c) *Design of uptakes and ports.* There is every advantage to be gained from having the gas and air ducts as straight as possible on their way to the regenerators, with smooth bends of large radius where changes of direction are necessary. In this way the "chimney pull" necessary to draw in the gas and air is kept low and, moreover, deposition of the heavy tarry vapours in the gas by direct impingement on the walls of the ducts is minimized. In the regenerators, of course, the brickwork lattice must obstruct the streamline motion of the gas and air to some extent, so as to cause some turbulence, if efficient heat interchange is to occur. In the furnace "ports" where the air and gas meet, and where combustion begins, turbulence should be encouraged otherwise the air and gas pass into the furnace in two streams one above the other, restricting combustion to the surface where the streams are in contact, and eventually forming a "blanket" of burnt gases effectively separating the still unburnt gas in the lower part of the gas stream from the unused air in the upper part of the stream of air. In some cases the gases leaving the furnace may have an appreciable proportion of unburnt gas in the lower layers, and over 20% excess air still available in the upper layers, so that combustion is still going on, wastefully and imperfectly, as mixing occurs later in the regenerators on the exhaust side. Here again the application of physical principles points to simple methods of ensuring that mixing occurs efficiently as the air and gas pass into the ports from the "uptakes" from the regenerators, and with such mixing increased efficiency is obtained by ensuring complete combustion with less air intake. The full calorific value of the fuel is thereby made use of in the furnace itself, and a higher temperature is developed because less of this heat is used in raising the temperature of an excessive amount of air.

(d) *Insulation and utilization of waste-heat.* Reference has been made to a method of assessing the thermal efficiency of a furnace. Obviously, since large amounts of heat are lost by conduction through the furnace walls and crown, some considerable increase of efficiency could be obtained by the proper use of heat insulation. It must be recognized, however, that heat must be actually abstracted from the glass as it passes from the melting end to the "working end", in order that the viscosity may increase to the extent necessary for the drawing, rolling or other shaping process to which the glass is eventually to be subjected. Insulation cannot therefore simply be applied uniformly over a whole furnace without completely altering the balance of temperature in the tank, and thus interfering with or actually preventing production. The application of insulation

must, in fact, be studied with full regard to the conditions which must be fulfilled in order that the glass may be melted economically and later be delivered at the working end in a condition ready for processing. The problems involved have by no means been fully explored but are being studied intensively; methods of abstracting heat in such a way that it can be used for preheating boiler feed-water, in low-pressure boilers, in hot water circulating systems for heating offices etc. are also being considered. Substantial progress has already been made, resulting in fuel economies as high as 20% in some cases, but there is obviously scope for much greater economies yet, in view of the very small proportion of the heat supplied which is actually used in melting glass.

Most of the foregoing section refers to tank furnaces only. Some work has been done also to improve the efficiency of pot furnaces. The problem of increasing the efficiency of pot furnaces is more complex owing to the necessity for heating the furnace and pots to melting temperature (*ca.* 1500° C.) each time the pots are refilled, then allowing the furnace to cool to working temperature (*ca.* 1000° C.) when the glass is fully melted and ready for working.

(e) *Furnace operation and control.* As has already been mentioned, there are obvious uses for pyrometric equipment in controlling furnaces, and there is obvious scope for the physicist in ensuring that the equipment is properly arranged and its indications properly translated. With such equipment properly installed, it may appear reasonable to assume that, if the rate of adding raw batch is maintained constant, satisfactory melting conditions will be assured, provided that a constant furnace temperature is maintained. It will be realized, therefore, that a distinctly baffling problem was presented when it was discovered that this assumption was not justified in practice. On occasion, with all temperature indications normal and rate of working constant, an experienced tank-worker would remark that the furnace was "going back," and within a comparatively short time it would become clearly apparent that melting was no longer proceeding satisfactorily.

In such cases it is worth while to go back to the general principle that the function of an instrument is to replace the skill and experience of a craftsman by means of a pointer moving over a scale. Obviously if the craftsman can detect something not indicated by the instruments, either the instruments themselves are not being properly used or they are not of the types necessary to measure all the things which the eye of the craftsman observes. Usually the craftsman can describe in his own language what he sees, and it is for the scientific and technical staff to translate that into physical quantities and to devise methods of measuring all of them.

To the skilled furnace-man the appearance of the flames in the tank is the first criterion by which he judges melting conditions; he also looks at the melting "lumps" of raw material floating on the glass, at the furnace walls, and so forth, but his main concern is with the appearance of the flames. If the flames are "thick", i.e. opaque, the furnace-man is satisfied, but a "thin", semi-transparent flame is, to him, an indication of poor melting conditions.

The difference in appearance is due to differences in the proportions of heavy tarry vapours in the gas; the quality of the flame should, therefore, be associated with differences of calorific value. This would be difficult to establish, however, because the methods of taking gas samples for determinations of calorific value are

known to be unsatisfactory, in that any heavy vapours present in the gas are practically certain to be condensed during sampling and so do not contribute to the calorific value as determined on the sample. The fact that the furnace temperature is fully maintained would, however, appear to indicate that, even if the calorific value of the gas is low, the supply of gas has been increased proportionately, so that some explanation other than an assumed change in calorific value must be sought.

Consideration of the way in which heat is transferred from the flames to the melting glass offers the solution to this, at one time inexplicable, effect. In general, the flames pass across the furnace without impinging on the surface of the glass, though they may lick the surfaces of the large "lumps" of raw batch material floating on the glass. The furnace crown is heated directly by the hot gases from the flame, though the flames do not play directly on the crown. Heat transfer to the floating raw material and to the glass is thus mainly due to radiation, part from the crown and part from the flame itself. If the temperature of the crown is kept constant, as indicated by the pyrometers embedded in the crown, the radiation from the crown on to the glass will be constant. Variations in luminosity of the flame will, however, cause variations in the heat radiated directly to the glass by the flame itself, and, although an "opaque" flame may partly screen the glass from the crown radiation, the total radiation received by the glass will be higher if the flame is strongly luminous. Evidently, therefore, some means of recording the luminosity of the flame should be capable of detecting the changes in flame quality on which the skilled furnace-man judges melting conditions. In experiments made with a "luminosity meter" directed horizontally across the flames this has actually been found to be the case. A fall in the luminosity was indicated by the luminosity meter each time the gas was described as "poor" or "thin", giving warning of poor melting conditions before any changes in tank temperature were indicated by the thermocouples of other pyrometric equipment used on or in the tank. So far as can be seen, therefore, the luminosity meter used in conjunction with the usual pyrometric equipment can be regarded as having completed the instrumentation needed to replace the eye of the skilled furnace-man.

(f) *Temperature distribution and convection currents.* Investigation of convection currents or circulation in pots and tanks, due to differences of temperature between different parts of the molten glass, has been carried out extensively, mainly in connection with the possibility of reducing the inhomogeneity of the glass produced. It has long been the practice to cause turbulent motion by mechanical stirring, when homogeneity is essential, as in making optical glass. The effectiveness of stirring depends, of course, on the thinning out of the layers of glass so as to increase the areas of contact between them, any differences of composition being adjusted as the result of interdiffusion between adjacent layers. In simple stream-line convection there is nothing to encourage the thinning out of the layers in this way, consequently differences of composition are adjusted only very slowly. Rate of diffusion can, of course, be increased by raising the temperature of the glass and it is the regular practice to maintain the molten glass at a comparatively high temperature for some time after melting is completed, in order that inhomogeneity may be reduced to a minimum. Attack on refractories, however, increases rapidly as the temperature is raised, and the risk of inhomogeneity being

produced by fresh refractory material being taken into solution sets a limit to the extent to which raising the temperature may be effective in promoting homogeneity of the melt. Some thinning of the layers of glass can be produced by causing the glass to flow through an aperture, below glass-level, in passing from the melting end to the working end. This is a very effective way of improving the glass from the point of view of uniformity of composition, as the glass enters the "dog-hole" or "throat" whilst still very hot, so that diffusion is still occurring at a high rate as the glass is being drawn through. Contamination by solution of the refractory blocks surrounding the dog-hole is, of course, a source of fresh inhomogeneity, but can be minimized by the use of specially resistant blocks for lining the dog-hole itself.

Methods have been tried for creating turbulent motion without recourse to mechanical stirring. The obvious line of attack is to cause local rise of temperature near the bottom of the glass, the most convenient method being to pass an alternating current through the glass between suitably disposed electrodes. Experimental work on these lines has given results of very considerable promise, though much development work is still needed before this method can be applied to large scale production.

Investigations into the directions and rates of flow of molten glass at various depths in tank furnaces, though revealing nothing that could not have been deduced from the ordinary laws of convection, yielded a fuller understanding of what actually happens in a glass-melting tank and enabled a close estimate to be made of the total "forward" flow. By comparing this with the amount of glass "drawn" per hour, the mean number of times any portion of the glass circulates up and down the tank could be deduced; the observed directions of movement near the side walls offered full explanation of the types and directions of the score-marks produced on the side-blocks and so forth. In considering the operation of a new tank incorporating new features of design, also, a knowledge of what may be described as "normal" circulation enabled certain faults in working to be identified almost at a glance, and correctly attributed to one of the new features, so that rectification could be proceeded with at once and with minimum loss of production.

The measurements of temperature at various depths have been of value in another way. If a tank is not deep enough the bottom becomes unduly hot, necessitating the use of high quality bottom blocks if corrosion is not to be excessive. If, on the other hand, the tank is unnecessarily deep the glass at the bottom becomes practically stagnant; this protects the bottom blocks against corrosion, but it also gives opportunity for devitrification to occur, i.e. for the growth of crystalline silicates in the almost motionless glass near the bottom. This, again, is of little importance so long as the crystals remain undisturbed; nevertheless it is always a potential danger, since any temporary change in melting conditions may cause some of this glass to be disturbed and so bring the crystals into the circulating glass and into the finished product, necessitating rejections and consequent loss of production, sometimes over periods of several days. The observations made on the variation of temperature with depth in different glasses showed the extent to which the temperature distribution is affected by the colour of the glass, thus enabling the most suitable depth to be determined for tanks in which different types and colours of glass are to be melted.

(g) *Devitrification studies.* The problem of devitrification must be faced whenever a new type of glass is being developed, and is never far in the background even with glasses of well-established composition. Determinations of "liquidus" temperatures above which no separation of crystals occurs, even in periods of 50 or 100 hours, have assisted considerably in deciding on modifications of composition to give a greater margin of safety against devitrification under production conditions. Studies have been made, also, with special microscopic equipment, to determine the changes of dimensions of individual crystals in glass held at various temperatures. These show that the rate of crystal growth is greatest over a comparatively narrow temperature range; above this the tendency for crystals to form is slight, while at lower temperatures the rapidly increasing viscosity of the glass reduces the rate of diffusion to such an extent that crystal growth is virtually impossible. Obviously, by arranging production conditions so that the glass in the working end is held a little above this range and is then cooled rapidly through it during working, risk of trouble from devitrification can be very considerably reduced.

(h) *Variation of viscosity with temperature, in relation to production processes.* For many production processes rapid "setting" of the glass in cooling through some particular temperature range is of advantage or may, indeed, be essential. In the modern method of making sheet glass, for example, the sheet is drawn vertically from a bath of molten glass and must become substantially solid within distance of some five to six feet from the surface from which it was drawn. This obviously requires a very considerable increase in viscosity over the temperature range through which the glass is cooled in rising through this short distance. In deciding on a suitable composition for sheet glass to be made by this process much valuable guidance was obtained from viscosity–temperature curves taken in the laboratory on experimental melts of different glasses, and changes of composition, resulting in improved rates of drawing, have been made as the result of further work of this type.

In making pressed articles there is not the same necessity for a quick-setting glass, since the pressed shape can be allowed a sufficiently long "dwell" in the mould for distortion to be avoidable when the shaped glass is removed. With a quick-setting glass, however, the pressed shape can be withdrawn after a comparatively short "dwell" without risk of collapse or deformation, and rates of production have been increased by between 20% and 50% in recent years, by the adoption of quick-setting glasses having compositions based on laboratory determinations of viscosity over the appropriate temperature ranges.

(i) *Strain distribution during processing—control of annealing processes.* The production of flat drawn sheet was delayed many years by failure to overcome the waisting effect which occurs in attempting to draw any thin sheet of liquid as, for example, a flat film of soap solution. The problem was eventually solved by making the edges thicker than the centre and chilling them so as to make them rigid, a remedy which appears obvious now but which took many years to discover and apply. The chilled edges are capable of holding out the central portion of the sheet to full width whilst it is soft and thus prevents the waist from developing. As the central portion becomes rigid, however, a system of strains is set up in the sheet, the chilled edges being put into compression with, of course, a compensating

tension over most of the rest of the sheet. The compression strains in the edges are a safeguard against cracks starting at the edges and breaking the sheet across. The tensions over the other parts of the sheet should, preferably, be uniform or rising smoothly to a flat maximum at the centre, though usually there is a reversal to compression over the central zone.

In operating the process it is essential to ensure that the stress distribution is as nearly symmetrical as possible, and that the strains are nowhere excessive. If, for example, the edge compression is unduly high the tensile strain at right angles to the edge may cause a portion of the edge zone to crack away, disturbing the balanced strain system across the sheet and tending to cause "cross-breaks" as the strain system readjusts itself. Similarly, if the distribution of strain is such as to cause a longitudinal compression in the central zone, the associated tension across the centre line may cause a "split-up" extending some distance up the middle of the sheet. The strains can be controlled by various means applied just above the surface of the molten glass in the drawing kiln, and by controlling the cooling of the glass after the sheet has become rigid. In general, observation of the strain distribution is unnecessary, but by means of "crossed" Polaroid sheets placed in front of and behind the sheet a rapid survey can be carried out whenever a check on the strain condition is desired, quantitative measurements being made by means of a Babinet compensator if necessary. Inspection through crossed Polaroids also shows up inclusions, whether opaque or clear, due to the strains set up in the glass round any "drop" or "stone" differing in composition from the main body of the glass.

A similar method of detecting defects due to inclusions or to "cords" (inhomogeneity) in the glass, as well as of imperfect annealing, can be used for rolled glasses or for articles of practically all kinds, the Polaroid sheets being mounted above and below or on opposite sides of the conveyor belt as the ware comes to the end of the annealing lehr. Observation of the condition of the finished glass is the most satisfactory method of controlling the operation of a lehr, the temperature distribution and the rate of travel of the article being adjusted until the annealing is shown to be satisfactory. When the conditions which give proper annealing have been established the corresponding lehr-wall and lehr-crown temperature may be taken as a guide to the quick readjustment of the lehr condition if for any reason production is interrupted, but it must be recognized that these temperatures will not be suitable if the type or size of article is changed, owing to the different amount of heat being brought into the kiln by the hot articles and the different rates of cooling of articles of different dimensions, particularly thickness.

The "annealing" referred to here is what may be described as "commercial strain-free" annealing; the problems of annealing optical glass are dealt with later (§ 4 (ii)).

(iii) *Other applications of physics in glass manufacture*

It would be possible to deal in this kind of way with practically every stage in the manufacture of glass and every detail of the manipulation processes to which it is subjected. The physics of glass cutting and glass grinding was discussed and described in an extraordinarily valuable paper by F. W. Preston in 1923. The flow of glass which occurs during polishing as predicted by Beilby, has been less

conclusively dealt with in published papers, but there is much evidence in support of these views and it would be difficult to account for the established facts on any other hypothesis. For observing quality of polish as regards the absence of scattering from a polished surface, visual observation has, so far, not given place to any instrumental method; but for detecting minute local departures from flatness a modified "Schlieren" method is capable of a sensitivity not normally obtainable except by interferometric methods. With such a method departures from flatness of less than $\lambda/4$ can be detected and recorded photographically.

Light transmission tests of glass, whether in the form of flat plates or in shapes of various forms, obviously come within the sphere of the physicist. For nominally "white" glass, transmissions must in many cases be not less than some specified minimum value, while for coloured glasses the density and hue must also fall within specified limits. Various standard methods of test have been developed for practically all observations of these types, including tests for infra-red and ultra-violet transmission as well as for the transmission of visible light.

Spectrophotometric observations are not normally used for test specifications, but such methods could be used in many cases as a guide to the chemist responsible for composition and batch mixture, and also as a guide to the furnace-man, when coloured glasses, sensitive to reducing or oxidizing conditions, are required to comply with specific requirements.

Control of toughening processes and the non-destructive testing of toughened articles are again matters for the physicist. Except when the article is of an unsuitable shape the toughening strains can be determined by directing a beam of polarized light edgewise through the glass, and determining the strain distribution through the thickness of the glass by means of a Babinet compensator. In general, uniformity of toughening of flat glass and of simple shapes can be determined by observing the pattern shown when the toughened glass is observed "flat-wise", or along directions at right angles to its surfaces, between crossed Polaroid sheets. The distribution of strain in toughened glass as revealed by the Babinet compensator shows remarkably close agreement with the type of distribution predicted some twenty years ago by the author, from mathematical investigation, despite the somewhat arbitrary and simplified assumptions made in the mathematical treatment.

And so the tale could be extended almost indefinitely; surface tension measurements have been made on glasses contaminated in various ways, to determine whether "drops" of glass contaminated in one way or another would tend to remain as spherical inclusions or to disperse rapidly into the surrounding glass. But to give merely a catalogue of such applications of physics would become dull and uninteresting; their range and extent can be judged, however, from the fact that the physicist has been consulted even on problems concerned with the construction of packing cases. Consultation in connection with improvements in the design of packing cases may not be regarded as inappropriate or startling; in one particular case, however, an investigation was carried out to determine the extra number of blows and, therefore, the extra time needed to drive nails of a particular type into cases made from wood not normally used, in order that fair piece-work rates might be established for the case-makers engaged in making those particular cases!

§4. THE CONTRIBUTION OF PHYSICS TO MORE FUNDAMENTAL KNOWLEDGE OF THE NATURE OF GLASS

(i) *Glass structure*

Up to 1925 or thereabouts our knowledge concerning the fundamental nature or structure of glass was limited to the fact that glass is an inorganic amorphous or non-crystalline isotropic solid or quasi-solid. According to the point of view taken it could equally well be regarded as a liquid of virtually infinite viscosity, or as a mixture of mutually inter-soluble inorganic materials which, as a result of rapid cooling, constituted a "super-cooled" solution or mixture of super-saturated solutions.

Investigations directed towards the obtaining of a more precise understanding of the nature of glass proved entirely unpromising until X-ray diffraction methods were tried by Warren and his co-workers in America. As a result of their work it was shown that any glass of the normal "silicate" type could be represented as an irregularly aggregated system of silica groupings, with the various alkali or other basic oxides distributed through the system, and the whole system linked together through the oxygen atoms. The structure of glass was thus shown to be built up from the same units as go to build up the three crystalline forms of silica: quartz, cristobalite and tridymite. In crystalline silica each silicon atom is linked with four oxygen atoms situated at the corners of a regular tetrahedron, with the silica atom at the centre of the tetrahedron; each oxygen atom is, of course, linked with two silicon atoms in adjacent tetrahedra. The arrangement of the tetrahedra is slightly irregular in a pure silica glass (fused silica), sufficiently so to destroy any crystalline properties in the ordinary sense. The mean spacing of the tetrahedra is only very slightly greater than the regular spacing in tridymite, and, of course, there must be minute portions of the glass, a few millimicrons in extent, within which the arrangement and spacing of the tetrahedra are entirely regular.

In ordinary silicate glasses the average spacing of the silica tetrahedra is less close than in a pure silica glass. This is necessary in order to make room for the alkalis and other basic oxide constituents and, in general, the mean spacing increases as the proportion of basic constituents is increased.

This new light on glass structure gives a new angle of approach to many of the problems concerning glass and its properties. If glass consists of a network of silica tetrahedra linked together, in the main, by oxygen atoms, the process of glass formation must be regarded as the opening-up of the oxygen linkages in the originally pure silica, so that the basic constituents may become associated or combined with the silica to form the resulting glass. When an oxygen linkage between two silica tetrahedra is momentarily disrupted, one side of the gap is left without any oxygen atom at all, while on the other side there is an oxygen with one of its valency bonds unsatisfied. If, now, a divalent basic oxide, such as CaO, enters the gap, the gap can be bridged again by the Si–O–Ca–O–Si linkage replacing the original Si–O–Si link. If, however, an alkali oxide such as Na_2O enters the gap, one of the Na atoms becomes attached to the unsatisfied oxygen atom on one side, while the remaining NaO group takes the place of the missing oxygen at the corner of the other tetrahedron. The linkages with the silicon atoms can now be represented as Si–O–Na and Na–O–Si, and as both sides of the gap end in univalent sodium atoms there is now a complete break in the structure.

From considerations of this type it becomes obvious that, since a glass is a massive assembly of interlinked silica tetrahedra, the formation of a true, massive, glass is impossible unless each silica tetrahedron is left with at least two corners unencumbered by alkali oxides. In other words, if there are not more than two sodium atoms associated with each silicon atom the structure can be built up indefinitely, i.e. the limiting composition of a true sodium-silicate glass is $Na_2O.SiO_2$. Experimental evidence completely bears out this deduction. If silica and sodium hydroxide or sodium carbonate are melted together in various proportions, with the Na_2O in excess of the proportion corresponding with the above composition the melt quickly separates into two liquids phases as it cools, one phase solidifying as a glass of the composition $Na_2O.SiO_2$, while the other phase contains all the Na_2O present in excess of that proportion.

The introduction of CaO, either in addition to the Na_2O or in partial substitution for Na_2O, does not necessarily give results of the same type. Provided that the CaO is not present in so large a proportion as to prevent complete melting, the resulting glass contains all the CaO plus a proportion of Na_2O; the limiting compositions possible in glasses of the type $xNa_2O.yCaO.SiO_2$, x being not greater than unity, have not apparently been determined but are under investigation.

The viscosity of a melt containing CaO in partial substitution for Na_2O increases rapidly with increase of the CaO content, as would be expected if, in accordance with the theoretical discussion given earlier, the CaO enters into the structure as a bridge-linkage replacing the simple direct oxygen linkages between the silica tetrahedra. The solubility of the glass in water is also greatly reduced by the substitution of CaO for some of the Na_2O, again as would be anticipated.

This view that in glass-making the silica structure is opened and absorbs the basic oxides reverses the usually accepted idea of the process involved in glass-making, according to which the silica is "dissolved" by the molten sodium oxide or carbonate. If a glass can be made with silica present in any proportions in excess of the amount represented by $Na_2O.SiO_2$ even up to pure SiO_2 (fused silica) there is certainly no question of the silica being dissolved in the sodium oxide during the process of glass formation. On the other hand, it is quite clear that silica becomes saturated with Na_2O at the composition $Na_2O.SiO_2$, any excess of Na_2O being thrown out of solution as the melt is cooled.

These deductions and their experimental verification may savour more of chemistry than of physics; they are, however, direct corollaries of the structure of glass first revealed by physical methods. In considering this structure it is necessary continually to switch from its physical to its chemical implications, and to visualize the physical properties which would be associated with any particular chemical bonding or combination which may appear to be probable. For example, if $Na_2O.SiO_2$ represents the limiting composition of a sodium silicate glass at the high soda content end of the scale, what should be the viscosity of a glass having this composition? Obviously the viscosity at any particular temperature will depend on the mean size of the groupings formed by the two oxygen linkages left free on each silica tetrahedron for building on to the adjacent silica tetrahedra. The rapid decrease of viscosity with temperature indicates that the groups break down into smaller groups comparatively rapidly as the temperature is raised, which would

further imply that rapid cooling from any high temperature should give a glass of lower specific gravity and lower refractive index than would be obtained if the glass were allowed to cool slowly. The very high viscosity of the glass at ordinary atmospheric temperatures implies, also, that the mean molecular groupings are very large. This raises the question as to whether groupings of such size could be formed if the only linkages in the glass are those afforded by the two oxygen atoms on each of the tetrahedra. If it could be shown, as appears possible, that the mean size of the groups, as deduced from viscosity determinations, is significantly larger than could be accounted for by these direct oxygen linkages, the question of "co-valent" bonding between the silica tetrahedra arises. Such bonding could occur by the formation of electron rings round the sodium atom, including spare (unused) electron pairs on the oxygen atoms, the sodium atoms thereby taking some part in the building up of the glass structure instead of producing complete discontinuities wherever they enter into combination.

A comparison between the viscosities of soda-silica glasses and their potash-silica counterparts would throw some light on the probability or otherwise of co-valent bonding having an appreciable effect in building up the molecular groupings. In fact, there is a very considerable field to be explored on these lines, involving both physical and chemical investigations, but all based on the fundamental structure as revealed by Warren's work.

Purely geometrical considerations of the structure give rise to a further wealth of problems needing investigation. The alkali atoms or ions must obviously be found some space within the network of silica tetrahedra which is the essential characteristic of glass structure. In the structure of tridymite there are vacant spaces of definitely known shapes and sizes. In a glass there must be interstices of a similar character which, though not all of one size or shape, have mean dimensions not greatly different from the dimensions of the vacant spaces in tridymite. It is reasonable to assume that the soda atoms in glass are mainly or entirely housed within these interstices when the number of soda atoms is sufficiently small. If, however, the number of soda atoms exceeds the number of spaces available, each extra sodium atom must be made room for by the opening up the silica structure. The sodium atoms will then be no longer so well enclosed but will be freed comparatively readily, either by water or other solvents, or by the application of a small potential gradient through the glass. Effects of this type can very readily be tested. It is unlikely that the properties will change suddenly at a particular soda content, since irregularities in the glass structure tend to make all properties change continuously, but comparatively sharp changes in electrical and other physical properties might reasonably be expected to occur within a small range of composition around the composition corresponding with the complete filling of the interstices by sodium atoms. Evidence of such changes has already been found.

Differences of behaviour may be expected, also, between soda-silica glasses and potash-silica glasses, owing to differences between the atomic dimensions of sodium and potassium. The potassium atom is of very nearly the same size as the interstices in tridymite, and so might be freed less easily than a sodium atom. Work on these lines has already given promise of valuable developments in the production

of glasses which can be melted comparatively easily but having electrical properties not hitherto obtainable except with high melting point glasses.

Other deductions from and consequences of glass structure as revealed by x-ray diffraction methods could be indicated, and it is fair to claim that practically all our fundamental knowledge of the nature of glass has resulted from that one physical discovery.

(ii) *Annealing of glass*

In the ordinary way the term "annealing" is understood to mean the removal of strain due to the glass not having been cooled sufficiently slowly, or due to the existence of temperature differences between different parts of the glass during cooling. Obviously, if a mass of glass is cooled under conditions such that one part is appreciably hotter than an adjoining portion, the cooler portion will have become substantially solid and unyielding whilst the hotter part is still capable of slight flow, and will have contracted appreciably after "setting" by the time the hotter part has become solid. At that particular moment there will be no strain in either of these two portions of the glass, but when both are cooled to ordinary temperatures the portion which solidified last will be left in tension and there will be a balancing compression strain in the adjoining portion which solidified first.

In optical glasses uniformity of refractive index throughout the glass is a first essential. It has been known for many years that a glass left in a strained condition through imperfect annealing is unsuitable for the making of lenses of the highest optical quality, and it was therefore the universal practice to anneal optical glasses very slowly and to test for perfection of annealing by observation of residual strain in the annealed blocks. With the development of lenses of larger and larger aperture, required to comply with stringent specifications concerning definition and general optical performance, freedom from appreciable strain was not always found to imply uniformity of refractive index, and considerable controversy arose between the makers of optical instruments and the manufacturers of optical glass. By interferometric methods it could be shown that the variations of refractive index in an annealed block were such as might be caused by the strain distributions revealed by polarized light, but the variations of refractive index were very many times greater than could be accounted for by the compressive and tensile strains corresponding with the double refraction effects observed. Mathematical investigations of the strains which would be produced in pieces of glass of various forms convenient for mathematical discussion showed that, in general, the maximum shear strains were of the same order of magnitude as the maximum compressive and tensile strains in the same regions, consequently some effect other than mere mechanical compression or tension must be involved to account for the variation of refractive index.

To pursue the whole story in detail would be far too lengthy a matter for this Report, but the final outcome of the investigations into the problem showed that the refractive index of any portion of a block of glass depends on its thermal history. To obtain uniformity of refractive index, therefore, annealing arrangements must be so devised as to ensure that all portions of the block of glass are subjected to the same temperature-time schedule, a condition which was, of course, originally achieved by annealing the blocks extremely slowly.

During the investigations carried out it was found that, with one type of glass, a single piece, when subjected to a particular annealing treatment, could be made to have a refractive index 0·002 higher than the refractive index of the same piece after it had been subjected to a not greatly different annealing process, the strains in both cases being small and of the same order of magnitude. Either refractive index could be reproduced at will, or intermediate values could be obtained by annealing according to schedules intermediate between the schedules giving the refractive index difference of 0·002.

The physical changes in the glass resulting from the different heat treatments can be conjectured, though it would be interesting to investigate the matter further in order to obtain fuller evidence. The indications are that the molecular groupings are of different sizes depending on the thermal treatments given, a higher degree of "polymerization" being produced in the one case than in the other. Such changes in the size of the groups may be due presumably to growth through the direct oxygen bonds between the silica tetrahedra, though the progressive development of co-valent bonds during the cooling of the glass must not be excluded from consideration.

The breakdown of the molecular groups in glass is indicated also by other phenomena. Most glasses expand fairly uniformly when heated over a considerable temperature range, starting from atmospheric temperatures. The coefficient of expansion begins to increase perceptibly, however, at some particular temperature depending on the composition of the glass, and within a comparatively short range may increase to some three times its value at lower temperatures. The effect is a strong indication that "dissociation" or "depolymerization" is beginning to occur, and merits further investigation from this point of view.

(iii) *Colours in glass*

Reference has already been made to the value of colorimetric observations as a guide to the works chemist and furnace-man on the oxidizing or reducing properties of a particular batch or particular furnace conditions. Spectrophotometric observations on glasses produced from various batch mixtures can, however, be of considerable interest from a purely scientific point of view. In dealing with ordinary solutions it is, of course, well known that various compounds can be arranged in order of their relative oxidizing or reducing properties, so that any particular compound will reduce those higher than it in the series, but will oxidize any which lie below it. Similar relations exist between various ingredients introduced into glasses or present as impurities in the batch materials, and the chemist or physical chemist desiring to study such "oxidation-reduction" potentials in glasses can do so satisfactorily only by obtaining evidence from the colours produced, preferably by actual identification of the absorption bands.

The composition of the base glass also appears to have an appreciable effect in promoting or inhibiting reduction of certain colouring agents; a study of the colours produced by colouring agents in different base glasses should, therefore, be of considerable assistance to the chemist in investigating the more fundamental chemical aspects of glass.

Direct physical measurements to obtain information on these problems present considerable experimental difficulty; preliminary results, however, fall

into line with what has been deduced from colour observations in relation to the chemical properties of the base glass and of the colouring agents, and physical methods of attacking these problems offer good promise.

The observation of absorption curves should also throw some light on questions concerning the state of combination of colouring ions or colouring agents in glasses; dispersed colloidal colouring agents might, for example, show no displacement of the absorption curve such as would probably occur in the absorption curves of colouring agents which are actually in combination with silica in the glass.

(iv) *Other physical properties of fundamental interest*

In the comparatively early stage of our present knowledge of the nature of glass it is obvious that the investigation of physical properties of many kinds may be of value in extending that knowledge.

Measurements of electrical conductivity of glass should reveal something about the way in which the various atoms, particularly the alkali atoms, are built into or enclosed within the general structure. Variations of electrical conductivity with temperature may also give evidence concerning the breakdown of the molecular groupings, and may even be capable of being correlated with viscosity measurements at temperatures high enough for direct measurement of viscosity to be made.

Measurements of the radiation emitted by glasses at high temperatures give an indication of the extent to which the ions of the lighter elements are freed as the temperature is raised. Absorption measurements of coloured glasses at high temperatures might also be of value in this respect and also in other ways. It is, in fact, abundantly clear that the field for applying physical observations to the elucidation of fundamental information concerning glass and its nature is a wide one, and it is practically virgin soil. It can scarcely be expected that all methods of cultivation will give bountiful yields, but enough is known of the nature of the soil to indicate with reasonable certainty that many methods offer promise of ample reward.

EVAPORATION IN NATURE

By H. L. PENMAN,
Rothamsted Experimental Station, Harpenden, Herts

§1. INTRODUCTION

EVAPORATION plays some part in most human activity directed toward the production of food, clothing, shelter and comfort, but whereas industrial operations usually have the process under control, in nature it is as uncontrollable as the weather. In many of its aspects the study of natural evaporation is part of meteorology—" the most desperate branch of that desperate science " (Symons, 1867)—but it is a broad subject overlapping the borders of several other fields of scientific interest. Evaporation from oceans and land surfaces plays a large part in the heat balances of both; the precipitation and re-evaporation of this water demands the attention of meteorologists, hydrologists and all attempting to grow crops; and along with other weather factors the cycle of rain and evaporation maintained over geological time plays an important part in the formation of soil types and their world distribution.

§2. HISTORICAL SURVEY

Aristotle posed a fundamental question : " Is wind or sun more important in promoting natural evaporation? " His own answer was in favour of wind because it carried away the vapour, and although this answer has rarely been challenged, the question remained open while ideas about the nature of vapours and of evaporation were vague and techniques for measuring relevant physical magnitudes were inadequate or non-existent.

The beginnings of a scientific study of evaporation are due to Dalton, who not only initiated much in the technique of evaporation measurements, but also provided the indispensable requisite of progress, clear ideas on the nature of the process. Famous names are associated with earlier ideas : Halley and Benjamin Franklin believed evaporation to be due to solution of water in air; Lavoisier regarded it as the combination of fire with water. Dalton (1801–2) stated clearly that the space above a water surface could contain only a limited amount of water vapour, the maximum partial pressure of which was dependent on temperature and independent of the presence of other gases or vapours, and showed experimentally that where the partial pressure was less than the maximum value at the surface of the water then evaporation would take place at a rate directly proportional to the partial pressure difference, the " constant " in the relation increasing with increasing wind-speed. He gave what appears to be the earliest estimate of annual evaporation from Great Britain, based on the difference between the total rainfall and total river discharge, and was one of the first to set up what is in effect a micro-catchment area for the study of local and seasonal variations in evaporation from soil surfaces. A cylinder with impermeable walls was filled with soil so that the measured rain falling on the surface could only be lost as evaporation from the surface or as drainage through the bottom :

the difference between rainfall and drainage was equated to evaporation. This particular technique was adopted in many parts of the world, and useful information on evaporation from bare and cropped soils has been obtained by its use. In one of his essays (1834) Dalton showed himself aware of the nature of evaporation as an energy change, discussing its importance in the regulation of fluctuations in atmospheric temperatures.

Nineteenth- and early twentieth-century developments were helped by the notable contributions of Daniell, Regnault and Glaisher to hygrometry, and during this period a large number of empirical formulae for evaporation from an open water surface were produced, most of these taking what American workers refer to as the " Dalton " form :

$$E = a(e_s - e_d)(1 + bu), \qquad \ldots\ldots(1)$$

where E is the evaporation in unit time, e_s and e_d are the saturation water vapour pressures at the liquid surface temperature and at the dew point temperature respectively, u is the horizontal wind velocity, and a and b are constants. Among several exceptions, Italian workers from Ragona (1867) to Visentini (1936) have preferred to use statistical correlations of evaporation rate and air temperature, relative humidity and wind speed. As evaporation studies increased in economic importance there was a rapid increase in experimental work, largely in the western and middle western states of the U.S.A., where agricultural, geological and hydrological organizations worked out their own techniques. At Fort Collins, Colorado, for instance, Carpenter (1889) provided the first of a useful series of observations which culminated in the classic open-air work of Rohwer (1931), to be discussed in detail below. For many years now the weather stations of these western states have been measuring evaporation from a standard open water pan as a normal meteorological element.

The rate of evaporation from open water is important in itself and also as a standard for comparative purposes. The development of the study of evaporation from bare soil and from cropped soil was slower because it was much more difficult. For the whole earth, including all kinds of surfaces, Howard (1837) suggested that as one was the source of the other, then, broadly, the annual evaporation and annual rainfall should be equal, and on this basis Daubrée (1847) estimated that the energy used in evaporation was about one-third of the heat received from the sun. For smaller mixed surfaces, such as catchment areas, the extension and improvement of gauging of rain and run-off led to the accumulation of information about annual evaporation from these areas; Kittredge (1938) was able to obtain sufficient data to draw a (rainfall-*minus*-run-off) map of the United States for use in studying the distribution of forest land and the types of trees in it. Lloyd (1936–42), in Britain, has attempted to correlate the " losses " with seasonal variation of weather elements and with geological formation (see § 4 (iv)). For small homogeneous areas the drain-gauge technique became popular; Kohnke, Dreibelbis and Davidson (1940) have surveyed these installations, providing a bibliography of nearly 500 references, but many of these drain-gauge experiments were poorly set out and results have little value. A great number were designed for the chemical purpose of examining the leaching action of percolating rainwater, so that physical interpretation of the results is not always easy. Up to 1870 the drain-gauges

were small and shallow, and were filled in with soil in a way that destroyed its natural structure. In that year Lawes and Gilbert (Lawes *et al.* 1881–2) at Rothamsted took a great step forward in establishing three drain-gauges in which undisturbed blocks of soil 1/1000 acre in area had impermeable retaining walls built round them and collecting funnels fitted underneath. These gauges, 20, 40 and 60 in. deep, are still in continuous use, and as a rain-gauge of the same area was set up alongside there are daily figures of rainfall and drainage available from 1870 to the present day. Discussion of these records forms a substantial part of the account of evaporation from bare soil.

Eser (1884) and Wollny (1888) carried out experiments on small filled-in gauges, examining the effect of soil structure, soil type, soil depth, nature of surface cover, fertilizer and manure content, on amounts and rates of percolation, and gave some consideration to the effects of external factors such as seasonal rainfall distribution, inclination of the surface, and intensity of solar radiation. They found, as others have since found, that the total amount of evaporation from bare soil is limited, first by the amount of water that the soil can hold against gravity, second by the tenacity with which water is held by the soil when drying has been in progress some time, and third by the reluctance of soil to transmit water even when the moisture content is little removed from saturation.

For forest studies the drain-gauge technique was inadequate, and although Ebermayer (1879) attempted to use large natural gauges, and a modified form of his technique was later developed by Russian workers (see e.g. Katchinsky, 1934), there are still few reliable data for forest transpiration.

The growth of agricultural science stimulated work on the water requirements of plants, including considerations of the way in which plants abstract water from the soil and transpire it to the atmosphere. Direct evaporation from bare soil, being of no benefit to plant growth, was regarded as wasted water, and many experiments were designed to see if cultivation operations could reduce it. King (1890–91), using a soil sampling technique to follow seasonal changes in the moisture content of the soil, satisfied himself that stirring of the surface soil reduced evaporation, but his summer results are inconclusive, although the spring effect may be significant. Since then there has grown a large volume of papers on cultivation experiments, reviewed by Pereira (1941), most of which show quite clearly that during the midsummer growing season, when moisture conservation is likely to be important, evaporation from bare soil is not reduced by cultivation operations, and that the sole benefit of these operations on crop yield is in control of weeds.

The movement of water from leaf to air was widely studied; Briggs and Shantz (1913–17), for instance, studied the diurnal change in transpiration rate as a function of sunshine, temperature, humidity and wind, and found that when the water supply was adequate the leaves transpired at a rate nearly the same as that of an open water surface of the same area. This behaviour of the leaf as though it were a piece of wet blotting paper had been considered by Brown and Escombe (1900) in a classic of biophysics. Theirs was the problem of assimilation—analytically the same—in which they sought to explain why a leaf with only 1% of its surface permeable to gases could absorb as much as $\frac{7}{12}$ the amount of CO_2 absorbed by a dish of caustic soda of the same area, exposed to the same

still atmosphere. With the backing of theoretical work of Stefan (1881), later reproduced by Jeffreys (1918), their rather crude experiments satisfied them that the rate of evaporation from a circular surface into still air is proportional to the radius and not to the area of the circle. (The theory is the same as that for the acoustical end correction of a cylindrical pipe.) More recent experimental work (Powell, 1942 ; Penman, unpublished) has shown that the evaporation rate is proportional to $d^{1\cdot4}$, where d is the diameter of the circular area, but in either case it is obvious that if no other factor intervenes, by choosing a high population of small enough holes the total flow through the holes could be greater, in theory, than from a continuous sheet. There is still a wide belief that the evaporation rate from a continuous surface of water can be increased by covering it with a thin impermeable sheet containing a large number of small perforations. The " other factor " is revealed by transposing the problem : why is the evaporation rate from the continuous surface relatively so small? The answer was given by Maskell (1928) who pointed out that for both continuous and perforated surfaces there is a diffusion resistance to flow from the surface as a whole and for the perforated surface there must be added to this the combined diffusion resistance of all the perforations in parallel. The experimental work on leaves showed, and, when properly interpreted, the theoretical work confirmed, that most freely transpiring leaves would behave very nearly as open water surfaces, and similar arguments would lead to a similar conclusion for saturated bare soil, since here too the evaporating surface is effectively discontinuous, the mineral components being completely impermeable. Research in plant physiology during the period established one fundamental point on which leaf and soil systems differ. The perforations in the leaf—the stomata—are the openings through which practically all gas and vapour movements occur and their opening is controlled primarily by light. For most plants, therefore, transpiration is limited to daylight hours so that, while comparisons of bare soil and open water evaporation need not be dependent upon season and latitude, comparison of crop transpiration and open water evaporation might have to include these factors.

One by-product of the growth of biological interest in transpiration was the introduction of artificial leaves consisting of porous surfaces—filter paper or earthenware—kept moist by movement of water under surface tension from a reservoir, the evaporation being measured by changes in weight of the whole system or by measuring volume changes of water in the reservoir. Many forms were developed and they were, and still are, widely used in spite of their erratic behaviour and the great difficulty of specifying what they are measuring. Shaw (1877, 1881) examined some of the more popular forms (Wild, Lamont, De La Rue and Piche) and reported unfavourably on their performance.

The main activity of the nineteenth century may be summarized as the collection of data. For open water surfaces there were many attempts to find formulae to fit the facts; for soils, bare and cropped, evidence accumulated in various ways depending on the particular interest of the experimenter. During recent years a tributary has joined the main stream. This represents the activity of the meteorologist, directed, like that of others, to determining the amounts and rates of evaporation, but emphasizing a new aspect, namely the way in which the vapour is taken up and the mechanism of its distribution in the atmosphere.

There is now a scientific interest all along the chain of events: the soil physicist is concerned with where the water comes from and how it moves in the soil, and shares with the biologist some of his desire to know how the plant takes it from the soil, how the water moves through the plant and the way in which it passes from the leaves; the meteorologist is concerned with the dispersal of the vapour, i.e. with what we may conveniently regard as the properties of the " sink ".

The analytical approach to the problem has been principally on an aerodynamic basis, with some attempts to make a physical approach on the basis of energy considerations. One of the first results of this work has been to show the need for still more experimental data, because, in spite of the great number of evaporation observations that have been reported, the contemporary meteorological data are rarely comprehensive enough. The present phase may be regarded as one of finding facts to fit formulae.

§ 3. THEORY OF EVAPORATION

(i) *Aerodynamic approach*

The study of air flow in wind tunnels (see Ewald *et al.* 1930, for a review) has led to ideas that can be carried over to open air studies. Where the flow over a flat surface is turbulent, the horizontal motion is characterized by zero velocity at the surface, a rapid increase in velocity in a thin layer next to the surface and then by a velocity distribution with height that has been expressed in two forms:

$$u - u_1 = \frac{1}{k_0}\sqrt{\tau_0/\rho}\,\log z/z_1 \qquad \ldots\ldots(2)$$

and

$$u/u_1 = (z/z_1)^{n(2-n)}, \qquad \ldots\ldots(3)$$

where u, u_1 are the horizontal velocities at heights z, z_1, k_0 is a constant (von Karman's constant), τ_0 is the shearing stress near the surface, ρ is the density of the air, and n is a constant depending on the vertical temperature gradient in the air ($\div\frac{1}{2}$ for normal conditions). Open air results fit form (2) (Best, 1935); wind tunnel results have given form (3). Over a ten-fold range of z the two are almost identical, but the subsequent divergences are so great that the two equations must not be regarded as alternative forms of the same expression.

The net vertical motion is zero, but at any instant, outside the thin layer of rapid shearing, there will be vertical motion of masses of air with instantaneous eddy velocities ω'.

When air with these velocity characteristics begins to move over a surface having a different thermodynamic potential, such as vapour pressure or temperature, a transfer of vapour or heat will begin. The first flow will be through the sublaminar boundary layer and the appropriate transfer coefficient will be the molecular diffusion coefficient or the thermal conductivity. Outside this thin layer, damper or warmer air will be carried away in the eddies characteristic of turbulent flow to mix with the main air stream. Primarily a transport of momentum, this mixing process and the resultant distribution of velocity with height will depend upon a coefficient of eddy viscosity, so that both vapour transfer and heat transfer will be largely governed by the same general law, the final details of mixing being dependent upon the corresponding molecular diffusion coefficients. Recent work, in which values of the eddy transport coefficient

and the relation between velocity and height appear as consequences of the theoretical analysis, has grown out of the original work of Jeffreys (1918) who imposed the conditions that outside the layer of rapid shearing the wind velocity and coefficient of eddy diffusion were independent of height. For a steady wind, mean velocity u, blowing along the x axis, and a coefficient of eddy diffusion, K, the transport equation becomes

$$u\frac{\partial V}{\partial x} = K\frac{\partial^2 V}{\partial z^2}, \qquad \ldots\ldots(4)$$

where V, generalized as a potential, is here the vapour content of the air, and the equation applies only outside the layer of rapid shearing. The rate of evaporation per unit area, E, is $-K\rho\partial V/\partial z$ and integration of (4) leads to

$$E = V_0\rho\sqrt{Ku/\pi x} \text{ per unit area}, \qquad \ldots\ldots(5)$$

where V_0 is the difference between V at the surface and at a great distance. The evaporation rate thus decreases as the distance, x, downwind increases. For a circular area, radius r, the *total* evaporation per unit time is

$$E = 3{\cdot}95V_0(Kur^3)^{1/2}, \qquad \ldots\ldots(6)$$

and for all surfaces the evaporation rate is proportional to $u^{1/2}$. A few points, relevant to the subsequent discussion, are to be noted. The value of V_0 is vague. Setting $V_0 = V_s - V_\infty$, the " surface " value is indefinite, for though it is usually accepted that V_s is the value at the evaporating surface, the analysis suggests that it should be the value just outside the thin sublaminar boundary layer, which may be very much smaller. The value " at a very great distance ", V_∞, is equally indefinite and may be unobservable. The eddy diffusivity is inserted as an atmospheric constant; there is no guide to its magnitude in the analysis itself and for this we must turn to the work of Prandtl (Ewald *et al.* 1930, and Taylor, 1922).

Prandtl assumes that an eddy originating at a height z will move with a vertical eddy velocity, ω', to a new level $z+l$, conserving its properties until it reaches the new level, and will there mix with its new environment. This " path of mixing", l, is analogous to the mean free path of a molecule moving with unchanged momentum between one collision and the next. The mixing process is assumed to be discontinuous.

The change in the horizontal velocity, identified with the horizontal eddy velocity, u', is $-l\partial u/\partial z$ and Prandtl assumes that the vertical eddy velocity, ω', is equal to this and writes down the shearing stress, τ, as

$$\tau = -\rho\overline{u'\omega'} = \rho l^2|\partial\bar{u}/\partial z|\partial\bar{u}/\partial z, \qquad \ldots\ldots(7)$$

using a form that automatically takes care of signs. By definition

$$\tau = K\rho\partial\bar{u}/\partial z, \qquad \ldots\ldots(8)$$

and, adopting von Kármán's argument from experience of flow in pipes, the mixing length is given by

$$l = k_0\frac{\partial\bar{u}}{\partial z}\Big/\frac{\partial^2\bar{u}}{\partial z^2}, \qquad \ldots\ldots(9)$$

where k_0 is a dimensionless quantity.

From equations (7), (8) and (9), and the assumption that the shearing stress near the surface is independent of z, the velocity: height distribution of (2) is obtained, and

$$K = k_0^2 \left(\frac{\partial \bar{u}}{\partial z}\right)^3 \Big/ \left(\frac{\partial^2 \bar{u}}{\partial z^2}\right)^2. \qquad \ldots\ldots(10)$$

A wide range of experimental data from wind tunnels has led to the value $k_0 \doteq 0{\cdot}4$.

The assumption of discontinuous mixing is not readily acceptable. Taylor (1922) has attempted to formulate a theory of diffusion by continuous movement, and his work has been extended and amplified by Sutton (1934) and Pasquill (1943). The eddy is regarded as losing its identity as it moves through and mixes with its environment, and using the vertical eddy velocity, ω', as the identifying property, it is assumed that the correlation coefficient between the value of ω' at time t and its value at time $t+\xi$ will decrease as ξ increases; the mixing length is taken to be the vertical distance the eddy must move before this correlation coefficient becomes negligibly small. In effect, this is an addition to, and not a replacement of, the Prandtl concepts; it gives an opportunity to eliminate τ_0 from equation (2), and hence from equation (10) where it is implicit. In Sutton's treatment the correlation coefficient, R_ξ, is postulated as

$$R_\xi = \left(\frac{\nu}{\nu + \omega_1^2 \xi}\right)^n, \qquad \ldots\ldots(11)$$

where ν is the kinematic viscosity of air, ω' is the vertical eddy velocity and n is a constant ($\doteq \frac{1}{4}$).

From this new standpoint an alternative form for the diffusivity (cf. (10)) is obtained:

$$K = \frac{\nu^n}{1-n}\left[\tfrac{1}{2}\pi k_0^2 \left(\frac{\partial \bar{u}}{\partial z}\right)^3 \left(\frac{\partial^2 \bar{u}}{\partial z^2}\right)^{-2}\right]^{1-n}, \qquad \ldots\ldots(12)$$

and, again assuming constant τ_0, the velocity: height distribution of (3) is obtained. By substitution in (12), an expression for K is derived in terms of atmospheric parameters that can be measured:

$$K = \frac{(\frac{1}{2}\pi k_0^2)^{1-n}(2-n)^{1-n} n^{1-n} \nu^n}{(1-n)(2-2n)^{2-2n}} (u_1 z_1)^{1-n} (z/z_1)^{(2-2n)(2-n)}. \qquad \ldots\ldots(13)$$

Sutton was thus able to generalize the earlier work of Jeffreys. Because K is not independent of z, the evaporation equation (4), must be written

$$\bar{u} \frac{\partial V}{\partial x} = \frac{\partial}{\partial z}\left(K \frac{\partial V}{\partial z}\right), \qquad \ldots\ldots(14)$$

and the solutions take the forms

$$E(x_0, y_0) = C \bar{u}_1^{(2-n)/(2+n)} x_0^{2/(2+n)} y_0 \qquad \ldots\ldots(15)$$

for the *total* evaporation from a strip of length x_0 downwind and width y_0 across wind; and

$$E(r) = C' \bar{u}_1^{(2-n)/(2+n)} r^{(4+n)/(2+n)} \qquad \ldots\ldots(16)$$

for the *total* evaporation from a circular area radius r, where C and C' include the vapour pressure difference and are independent of $\bar{u}_1$, x_0 and r.

Pasquill sets out the somewhat complicated links between equations (13) and (15) and (16), reducing the variable parts of C and C' to

$$C \text{ (or } C') \propto \chi_0 \nu^{2n/(2+n)} k_0^{4(1-n)/(2+n)} z_1^{-n/(2+n)}, \qquad \ldots\ldots(17)$$

and gives a table of values of C and C' for values of $\chi_0=1$ gm/c.c., $\nu=0{\cdot}147$ cm^2/sec., $k_0=0{\cdot}4$, and $z=1$ cm. in the range $0{\cdot}20<n<0{\cdot}30$, where χ_0 is the difference between the density of vapour at the evaporation surface and that at "a great distance". As in Jeffrey's case the "surface" value is not precise, but the "great distance" may be regarded as the region of air to which the effects of the evaporation process have not penetrated; near the "leading edge" of the system this will be close to the surface and observable, but a few miles downwind it may be several hundred metres up and not normally accessible to observation.

From equations (15) and (16) the dependence of evaporation rate on wind speed takes the form $E \propto u_1^{(2-n)/(2+n)}$, and for a normal value of $n=\frac{1}{4}$, this becomes $E \propto u^{0{\cdot}78}$. The validity of this derivation is closely linked to the validity of the parallel derivation $u/u_1=(z/z_1)^{n/(2-n)}$, and this simple power law cannot be generally accepted. So far, mathematical analysis has only been possible, and then with great difficulty, on the basis of assumptions that contain this power law explicitly or implicitly.

Pasquill (1943), guided by the results of experiments in wind tunnels on various liquids in which he found inconsistent differences between observed and estimated evaporation rates, suggested that in diffusion problems the kinematic viscosity should be replaced by the molecular diffusion coefficient and, similarly, by the thermal diffusivity in heat transfer problems. In this way some account of the molecular transport phenomena was included, and this generalized treatment gave more consistent results.

American work has followed similar lines, but has employed the logarithmic law of wind change with height. Rossby and Montgomery (Brunt, 1939) have used the law in the form

$$u = \frac{1}{k_0}\sqrt{\frac{\tau}{\rho}} \log \frac{z+z_0}{z_0}, \qquad \ldots\ldots(18)$$

assuming the mixing length at $z=0$ has a finite value, z_0, which depends upon surface roughness, where z_0 may be about 1/30 of the height of surface irregularities (see Bagnold (1941) for an application of this to wind movement over sand). Typical values of z_0 are: short grass, 0·54 cm.; open grassland, 3·2 cm.; sea with slight swell, 3·9 cm. Sverdrup (1945), however, quotes $z_0=0{\cdot}6$ cm. for moderate wind velocities over the sea.

On this basis, Thornthwaite and Holzman (1939) have derived a flow equation using data obtained at two heights, z_1 and z_2. The result is

$$E = k_0^2(\chi_1-\chi_2)(u_2-u_1)/(\log_e z_2/z_1)^2 \text{ gm/sec/cm}^2, \qquad \ldots\ldots(19)$$

where all symbols have the same meanings as before. This form of equation has the great advantage of eliminating uncertainties about surface values. Note the direct proportionality between E and the difference of wind velocities.

A similar dependence of E on wind velocity has been obtained by Frost (1946). The analogy between eddy movement and molecular movement is rejected as inaccurate and non-essential to the theory. The mixing length is defined in a new way that takes into account the roughness of the surface, but the manner of introducing z_0 is such that it bears a relation to surface irregularities different from that of equation (18). Equations (7) and (8) are accepted, and as a working

hypothesis a new equation is introduced:

$$l = z^{(1-m)} z_0^{m}, \qquad \ldots\ldots(20)$$

where m is a non-dimensional constant equal to Sutton's $n/(2-n)$. Assuming τ/ρ is constant, equation (3) is obtained for the velocity: height law and

$$K = l^2 \partial u/\partial z \qquad \text{(from (7) and (8))}$$

$$= m u_1 z_1^{-m} z_0^{2m} z^{(1-m)}. \qquad \ldots\ldots(21)$$

This value of K is substituted in (14), and using the appropriate value of $\bar{u}$, the equation is solved using a new variable

$$\theta = z^{(2m+1)}/[(2m+1)^2 m z_0^{2m} x]. \qquad \ldots\ldots(22)$$

The result is

$$V/V_0 = \frac{\int_\theta^\infty \theta^{-(m+1)/(2m+1)} e^{-\theta} d\theta}{\Gamma\{m/(2m+1)\}}, \qquad \ldots\ldots(23)$$

giving the variation of vapour content with height as a function of m, x, z and z_0. Using $m = 1/7$, a table of values of V/V_0 is given for varying values of θ. The instantaneous rate of evaporation at x_0 per unit area is

$$E = \underset{z \to 0}{Lt}[-K \partial V/\partial z]$$

$$= \frac{(2m+1)^{1/(2m+1)} x_0^{-m/(2m+1)} (m z_0^{2m})^{(m+1)/(2m+1)} u_1 z_1^{-m} V_0}{\Gamma\{m/(2m+1)\}}, \qquad \ldots\ldots(24)$$

and, integrating, the *total* rate of evaporation is

$$E(x_0, y_0) = \frac{(2m+1)^{(2m+2)/(2m+1)} (m x_0 z_0^{2m})^{(m+1)/(2m+1)} u_1 z_1^{-m} y_0 V_0}{(m+1)\Gamma\{m/(2m+1)\}}. \qquad \ldots\ldots(25)$$

For $m = 1/7$ and $z_0 = 1$ cm., this reduces to

$$E(x_0, y_0) = 2{\cdot}85 \times 10^{-2} x_0^{8/9} y_0 V_0 u_1 z_1^{-1/7}. \qquad \ldots\ldots(26)$$

From (24) and an approximation of (23), valid when θ is small, x is eliminated, and the instantaneous rate is reduced to

$$E = m^2 z_0^{2m} u_1 (\chi_0 - \chi_1)/z_1^{2m}, \qquad \ldots\ldots(27)$$

where V is now replaced by the vapour concentration. Taking $m = 1/7, z_0 = 1$ cm., assuming the standard height of observation is 2 metres, and replacing concentrations by partial pressures, the daily rate at any distance satisfying the condition of small θ is

$$E = 0{\cdot}20\, u_2 (e_s - e_d) \text{ mm/day}, \qquad \ldots\ldots(28)$$

where u_2 is in miles per hour and e in mm. Hg.

The treatment is regarded as being applicable to moderate to strong winds.

Throughout this work there is a primary emphasis on the determination of a conductance, and the corresponding potential difference needed to deduce the flow rate is assumed to be easily measured. For completeness, a treatment in which the difficulties are reversed should be included. Movement through the thin sublaminar boundary layer is presumed to be governed by the molecular diffusion coefficient which is a known gas constant. The eddy diffusivity outside the layer is so very much greater than the molecular constant that it is reasonable

as a first approximation, to assume that practically all the vapour pressure drop takes place across the boundary layer and the gradient is determinable if the effective thickness is known. Brunt (1934) has given a value of this thickness based on Taylor's estimate of the skin friction at the ground as $0{\cdot}004\,\rho u_0^2$ per unit area. The shearing force per unit area is $\rho\nu u_\delta/\delta$, where δ is the thickness of the boundary layer, and equating this to the skin friction

$$\rho\nu u_\delta/\delta = 0{\cdot}004\rho u_\delta^2,$$

or, putting $\nu = 0{\cdot}15$,

$$\delta = 37/u_\delta \text{ cm.}, \qquad \ldots\ldots(29)$$

where u_δ is in cm/sec. In this u_δ is unobservable, but arguing from wind-tunnel experience Brunt sets $u_\delta = 0{\cdot}6\ u_\infty$ where u_∞ is the velocity at a great distance. Substituting in the diffusion equation the mean daily evaporation rate becomes (Penman, 1940)

$$E \doteqdot 0{\cdot}017(e_s - e_d)\bar{B} \text{ in./day}, \qquad \ldots\ldots(30)$$

where e_s and e_d are in millimetres of mercury, and B is the Beaufort wind force ($B = 2$ at $u \doteqdot 240$ cm/sec.).

Although it confers a false air of precision on a crude estimate, (30) can be turned into the same form as (28) to give

$$E = 0{\cdot}22\ u_2(e_s - e_d) \text{ mm/day}. \qquad \ldots\ldots(31)$$

As all these estimates of evaporation rates are measures of the ability of the atmosphere to take up further water vapour, i.e. to act as a sink for water vapour, they may be conveniently described as " sink strength " estimates.

(ii) *Energy balance*

The heat budget of various parts of the earth and its atmosphere have received considerable attention, but while oceanographers (Sverdrup, 1945) have made much use of an energy balance sheet to assess evaporation from the oceans, in the meteorology of the solid earth the position has tended to be that figures for evaporation were needed to enable other unknowns to be evaluated.

There are three phases of the problem. First there is an incoming supply of short-wave energy from sun and sky, measurement of which is relatively easy. Part of this will be reflected, the reflection coefficient depending upon the nature of the surface and the angle of incidence of the light, i.e. will include factors of season and latitude. Second there is an exchange of long-wave radiation between earth and sky, difficult to measure and not completely accounted for on theoretical grounds, so that one is dependent upon empirical correlations to assess the effects of water vapour and clouds on the exchange. Finally there is the problem of the fate of the net radiant energy available at the surface. The radiation intensity is conveniently reduced to an evaporation equivalent, the unit being 59 cal/cm^2/day. This is the amount of energy to evaporate $\frac{1}{10}$ gm. water/cm^2/day at 10° c., i.e. to evaporate 1 mm. water/day. If R_c is the incoming radiation, of which fraction r is reflected, the net short-wave income is $R_c(1-r)$. [For crops an additional factor to take account of energy used in photosynthesis may be ignored as it amounts to only $\frac{1}{2}$% or less.] The net long wave exchange has been estimated in two ways.

In the absence of any atmosphere the earth would radiate effectively as a black body. In a cloudless atmosphere the air itself radiates outward and downward, the radiation intensity being largely determined by its water vapour content. Ångström (1920), quoted by Sverdrup, and Brunt (1939) have given slightly different expressions for the net back radiation to a cloudless sky; Brunt's, the smaller of the two, is

$$R_0 = \sigma T_a^4 (1 - 0{\cdot}44 - 0{\cdot}092\sqrt{e_d}), \qquad \ldots\ldots(32)$$

where σT_a^4 is the Stefan radiation at mean air temperature and e_d is the vapour pressure of the air in millimetres of mercury.

Clouds appear to act as though they intercepted both upward and downward components of R_0, and if $m/10$ of the sky is covered by cloud, the back radiation, R_m, is given by

$$R_m = R_0(1 - \alpha m), \qquad \ldots\ldots(33)$$

where α depends upon cloud type. For high cloud (Cirrus) $\alpha \doteq 0{\cdot}025$, for medium height (Alto), $\alpha \doteq 0{\cdot}06$, and for low cloud (Cumulus), $\alpha \doteq 0{\cdot}09$ (Sverdrup). Using the last as an example, the heat budget, H, becomes

$$H = R_c(1 - r) - \sigma T_a^4 (0{\cdot}56 - 0{\cdot}092\sqrt{e_d})(1 - 0{\cdot}09m). \qquad \ldots\ldots(34)$$

Cummings and Richardson (1927) have adopted a simple treatment to avoid the difficulties of (32) and (33). For two similar surfaces, side by side, the total incoming radiation (short and long) will be the same for both, and equal to I say. The outgoing radiation is, implicitly, assumed to be given by $\sigma T_s^4 + c$ (where c is a constant) and will not be the same for both if their surface temperatures differ. Setting B for the back radiation

$$H_1 = I - B_1;\ H_2 = I - B_2$$

and

$$H_1 - H_2 = B_2 - B_1$$

$$= (dB/dT) \,.\, \Delta T = 4\sigma T_s^3 \Delta T,$$

or

$$H_1 = H_2 + 4(T_2 - T_1)\sigma T_1^3; \qquad \ldots\ldots(35)$$

i.e. if the surface temperatures are known and the heat budget for one surface can be measured, the value of H_1 for the other surface can be estimated. Their object was to calculate lake evaporation from small pan measurements nearby, and Cummings and Richardson were well aware that their research amounted to little more than a comparison of evaporation from two vessels of different dimensions. The next stage is common to this and the preceding more detailed treatment, namely the determination of evaporation when H is known.

Using the convenient notation of Cummings and Richardson, the available energy is used up in evaporation, E, in warming the air, K, in warming the water in the test pan, S, and in warming the surround of the test pan, C,

$$H = E + K + S + C. \qquad \ldots\ldots(36)$$

The change in the stored heat, S, and the amount lost by conduction, C, through the walls of the pan have to be measured unless conditions are such that these quantities are negligibly small compared with E and K; under natural conditions, this is often true if the period is reasonably long. The heat transferred to the

air from the surface, K, is transported by essentially the same mechanism as the vapour transfer, so that without knowing the nature of the mechanism, the ratio of K to E can be obtained (Bowen, 1926). In essentials, this ratio, symbolized by Bowen as R, and here by β, is obtained from one of the standard derivations of the wet and dry bulb hygrometer equation:

$$K/E=\beta=0{\cdot}27(T_s-T_a)/(e_s-e_d),\ ^\circ\text{F.}, \text{ mm. Hg.} \qquad \ldots\ldots(37)$$

and hence

$$E=(H-S-C)/(1+\beta). \qquad \ldots\ldots(38)$$

§ 4. EVAPORATION MEASUREMENTS

(i) *Open water*

A very complete bibliography of the subject was provided by Mrs. Livingston (1908–9), references to most of the important later papers are given by Powell and Griffiths (1935), Rohwer (1931), and Pasquill (1943), and a collection of evaporation records from all over the world was published by Follansbee (1934). There are rarely sufficient contemporary meteorological data supplied to enable one to test theoretical formulae.

The energy approach has been successfully applied to the oceans (Sverdrup). Direct measurements are difficult, but the following table shows the mean annual values for the Atlantic; it will be useful as a guide to world distribution of evaporation as the means for all oceans do not differ very much from those tabled. The estimated evaporation is based on (38).

The annual variation shows a midsummer minimum of zero with the main maximum from September–November, due to the small changes in ocean surface temperature. Similar winter maxima in evaporation from large areas of water have been reported by Folse (1929) for the Great Lakes.

Cummings and Richardson applied their treatment to estimate hourly evaporation rates from a tank of 520 sq. metres in area at Fort Collins, Colorado, and obtained good agreement with observed values. Cummings (1940) extended the treatment to lake evaporation, but had no experimental check on the calculations. Holzman (1941), in a general discussion of this work, concludes that the method is known to be precise, but it is difficult to measure all the parameters in the equations. In an empirical correlation of evaporation and energy (at Canberra, Australia), Prescott (1940) found that for a 3 ft. diameter tank the mean monthly value of E/R_c varied from 0·63 in February to 0·27 in July with an annual mean of 0·43. The seasonal change was related to the saturation deficit of the air:

$$\log(1-E/R_c)=-0{\cdot}044(e_a-e_d)-0{\cdot}065, \qquad \ldots\ldots(39)$$

where e_a and e_d are the saturation vapour pressures at air temperature and dewpoint temperature respectively.

The sink strength approach to evaporation has produced many empirical formulae. Using the normal value of n in (15) ($n=\frac{1}{2}$) we should expect the rate per unit area at distance x_0 downwind to take the form

$$E\propto(e_s-e_d)u_1^{0{\cdot}78}x_0^{-0{\cdot}11}, \qquad \ldots\ldots(40)$$

while for a circular area the average rate over the whole area (16) should be

$$E\propto(e_s-e_d)u_1^{0{\cdot}78}r^{-0{\cdot}11}. \qquad \ldots\ldots(41)$$

There are not sufficient open air data available to check the " constants " implicit in (40) and (41), and present evidence can only be used to test the factors individually. For wind tunnel work, Powell, Rohwer and many others have confirmed the dependence on $(e_s - e_d)$, and this is regarded as being so firmly established that it is usually the dependence of evaporation rate per unit partial pressure difference that is examined as a function of wind speed and surface dimensions. Powell (1942) reduced much information from other workers to a single diagram showing the index of u_1 in (40) to be 0·77 and of x_0 to be $-0{\cdot}22$; for (41) the corresponding indices were 0·56 (u_1) and $-0{\cdot}44$ (r). Pasquill, as already noted, found good agreement between theory and observation for a number of liquids, again in wind tunnel experiments.

Out of doors the position is not so clear. Frost has applied his equations to large scale oceanic evaporation. Taking $z_0 = 1$ cm. and $x_0 = 5000$ km. for the Atlantic, calculations of the vapour concentration in the air over the west coasts of the British Isles agree very closely with observed values (cf. (23)), and the estimated average annual evaporation between latitudes 50° and 55° N. is 68 cm. (cf. table 1). A long succession of workers have repeatedly obtained a relation

Table 1. Mean annual evaporation for the Atlantic. (After Wüst, 1922.)

Latitude North	Evaporation (cm./year)		Latitude South	Evaporation (cm./year)	
	Observed	Estimated		Observed	Estimated
50	—	62	10	143	130
40	94	94	20	132	145
30	121	121	30	116	140
20	149	135	40	81	100
10	132	142	50	43	65
0	116	125	—	—	—

between evaporation and wind velocity in the form of equation (1), in which the extrapolated evaporation rate at zero wind velocity is not zero. Several of them—Fitzgerald (1886), Carpenter (1889), Bigelow (1908), Meyer (1915), Carrier (1921), Freeman (1926), Lugeon (1928), Rohwer (1931), Banerji and Wadia (1932)—have obtained values of a lying between 0·3 and 0·5, although their values of b vary widely. Rohwer's work, most comprehensive of all, was carried out at Fort Collins, 5000 ft. above sea-level, using a tank 3 ft. square. His equation, based primarily on indoor experiments and checked out of doors, is

$$E = 0{\cdot}44\,(1 + 0{\cdot}27\,u_0)(e_s - e_d) \text{ mm/day}, \qquad \ldots\ldots(42)$$

where u_0 is the wind velocity at the surface (miles per hour) and the vapour pressures are in millimetres of mercury.

As zero wind velocity is never a condition of out-door work, it is possible in most cases to fit the experimental results with a curve that passes through the origin. Banerji and Wadia found that their data would fit Jeffrey's form both in velocity and in the effect of dimensions; Leather (1913), also working in India, found the velocity index to be 0·36. In general, however, such fitting is a forced

fitting, and difficult as it may be to accept it on theoretical grounds, the weight of existing experimental evidence is in favour of the form of equation (1).

Testing the dependence of evaporation rates on dimensions presents difficulties of several kinds. One is the obvious problem of building large watertight systems or finding existing man-made or natural systems on which reliable measurements can be made. Another concerns the effect of the surrounding surface and its behaviour. Where a test surface is set in the middle of an infinite sheet of dry non-hygroscopic material there is no doubt about the position of the leading edge of the system; where it is in the middle of a vigorously growing and transpiring vegetation which will act very much like an open water surface, at least during daylight, where is one to place the leading edge? This is the more normal condition for most meteorological stations and it will often happen that local exposure may be such that the hypothetical leading edge is at different distances for different wind directions, i.e. x_0 in (15), (17) and (40) is a variable unknown! Finally, the constant, k_0, in these equations is a mixing parameter based on wind-tunnel results and, as the scale of eddying is so very much greater out of doors, it is conceivable that the atmospheric k_0 may be greater than the wind-tunnel value. Thus the effective eddy diffusivity might increase as the scale of the evaporation surface increases.

Most of the available comparisons of pans of different sizes (Rohwer, 1934) were made to enable estimates of lake evaporation to be made from small areas nearby. From Rohwer's paper it appears that the depth of the pan is unimportant over 12 in., the shape is not critical (square and circular pans of the same size agree within a few per cent), pans 3 ft. square seem to give much the same evaporation whether sunk in the earth or floating on a lake, beyond 12 ft. diameter the dimensions seem to be unimportant, and maximum evaporation takes place from a pan raised above ground level on a timber grillage. The most frequent comparison has been with a sunk pan 3 ft. square, and in figure 1 is a summary of these com-

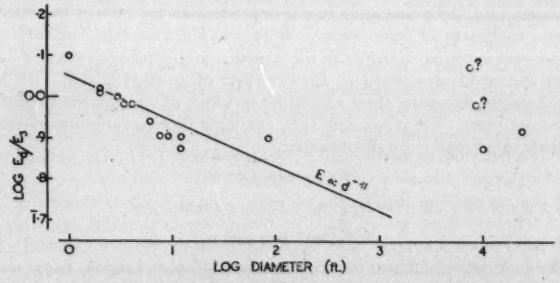


Figure 1. Evaporation/unit area and diameter of surface based on Rohwer (1934).

parisons. Over the diameter range 1–12 ft. most of the data are due to Sleight (1917), the values for 86 ft. diameter and a reservoir of 1800 acres are due to Rohwer (1931) and the largest surfaces were studied by Lippincott (1927) and

Harding (1922). For all these large surfaces the diameter of the equivalent circular area has been used as abscissa. The straight line in the figure has a slope of $-0{\cdot}11$ (equation (41)).

It appears that the nature of the surround is not very critical and it may well be that local eddies produced by the rim of the evaporation tank are much more important. It is not practicable to keep a tank always brimful. The effect of dimensions is not maintained throughout the range, and on the available evidence it appears that a tank 6–12 ft. diameter is an adequate representative of much larger areas.

Examination of Rohwer's earlier (1931) paper shows, however, that these conclusions are not quite so clear cut as they seem. He built a reservoir 86 ft. diameter in a grass surround and over a period of three years measured daily evaporation, mean wind velocity and relevant surface and air temperatures; using equation (42) he obtained estimates of the daily evaporation. From nearly 500 sets of observations the overall mean observed rate was less than the calculated rate and Rohwer took the ratio, 0·77, as the factor to be used to correct 3 ft. square land-pan measurements to an infinite water surface. These means are for all wind speeds; his average wind speed was 1·5 m.p.h., and only on very few days did it exceed 5 m.p.h. Examination of the daily records shows that on days when the wind speed was 3 m.p.h., or more, the correction factor was about unity, i.e. except at low wind speeds, where anemometer performance is uncertain the evaporation from a large water surface in a turf surround can be estimated from a formula based on a small surface in a similar environment.

Rohwer's very useful work gave the first data on another important aspect of evaporation. Many workers (e.g. Davydov, 1936) have assumed that as the coefficient of molecular diffusion is inversely proportional to the atmospheric pressure then the evaporation rate must be similarly corrected for pressure, and many empirical formulae have had a correction applied. This ignores other factors in the process. For example, in the crude boundary layer theory, leading to equation (30), the evaporation rate is dependent on the ratio of the molecular diffusion coefficient of water vapour in air and the kinematic viscosity of air. The pressure-sensitive variables in the equation are diffusion coefficient and air density and occur as a product; the net effect of pressure is zero. To a first approximation, therefore, there should be no effect of atmospheric pressure on evaporation rates. From equation (13), assuming there is no pressure-sensitive parameter other than ν, we should expect

$$E_p/E_{p_0} = (p/p_0)^{-n},$$

and if p is not very different from p_0, this can be reduced to

$$E_p/E_{p_0} = (1+n) - np/p_0. \qquad \ldots\ldots(43)$$

Rohwer took his equipment to six stations at heights ranging from -68 to 14 100 ft. above sea-level, applied equation (42) to observations of wind and temperatures, and from the ratio of observed to calculated evaporation obtained the pressure correction in the form

$$E_p/E_{p_0} = 1{\cdot}465 - 0{\cdot}465p/p_0. \qquad \ldots\ldots(44)$$

The treatment includes the effect of temperature as well as pressure and, since

the temperature changes would act in the same sense, the apparent value of n is slightly less than 0·465. As Rohwer's p_0 was 625 mm., equation (42), adjusted to 760 mm. total pressure, becomes

$$E = 0{\cdot}395(1 + 0{\cdot}27u_0)(e_s - e_d) \text{ mm/day.} \qquad \ldots\ldots(45)$$

Raman and Satakopan (1935) have applied Rohwer's equation to estimate probable evaporation rates from open water surfaces in India, assuming that the mean surface temperature would be the same as the mean air temperature. Evaporation maps, monthly and annual, are given for the whole of India, and areas are shown in which the estimated annual evaporation exceeds 150 inches. The energy content of the solar radiation that would reach these areas through a completely transparent atmosphere is equivalent to 196 inches of evaporation annually (Angot, 1883), and as only three-quarters of this energy survives scattering and absorption processes, even in the absence of clouds, the estimated evaporation must be too great.

There are not yet sufficient data available to permit plotting an evaporation map of the world, but a coarse indication of world distribution can be obtained from the tables of Follansbee (1934). Suitable correction factors were applied to results from small surfaces to give values expected to be representative of extended areas of open water. Annual totals range from 15 inches (in Ontario and Sweden) to 60 inches (in tropical and sub-tropical regions), in much the same way as the oceanic data of table 1. The annual variation of evaporation rate at land stations is very nearly in phase with the solar radiation cycle, and for Camden Square, London, the mean monthly values lie very nearly on a pure sine curve (Bilham, 1938).

(ii) *Bare soil*

Except in special circumstances the drying process reduces the amount of evaporable water and its maintenance depends more on soil properties and less on weather conditions as this amount decreases. Evaporation from bare soil is almost entirely an empirical study and in this review of the literature it will be convenient to deal first with general principles which are common to most soils and climates and then to consider the modifications arising from soil type.

Most permeable soils, if brought to complete saturation (i.e. no air content) and then allowed to drain will, in the absence of evaporation, tend to come to a fairly consistent moisture content known as " field moisture capacity ", conveniently abbreviated to " field capacity ". Subsequent evaporation will reduce the moisture content below field capacity and in this state rain or irrigation water equal in amount to the " deficit " must be added before drainage will begin. The main problems in evaporation studies of bare soil are the growth of the deficit during rain-free weather, and the variations of the aggregate evaporation during periods when alternate wetting and drying of the surface cause the deficit to fluctuate. Penman and Schofield (1941) have reviewed the records and the literature on the Rothamsted drain-gauges, and a summary of their conclusions follows. Annual evaporation from bare soil is very nearly constant; it increases slightly with increasing rainfall. Summer evaporation (April–September) is about two to three times as great as winter evaporation (October–March); the former is very dependent upon rainfall whereas the latter is independent of

rainfall (figure 2). The deficit rarely exceeds 0·7 in. of water and most of the summer evaporation occurs soon after the end of rainfall, and hence the total summer evaporation increases with the frequency of re-wetting, i.e. with total

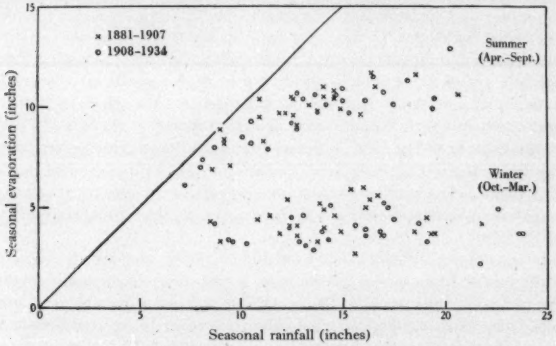


Figure 2. Summer and winter evaporation from fallow soil (Rothamsted).

summer rainfall. No other weather factor has any significant effect in summer (Sahni, 1941). During the winter, conditions at the site are normally characterized by wet surface soil, and atmospheric factors, other than rainfall, assume some importance.

Assuming general equality of mean surface and mean air temperature, later shown to be reasonable for winter periods (Penman, 1943), estimates of periodic evaporation rates based on (30) were made (Penman, 1940). Good general

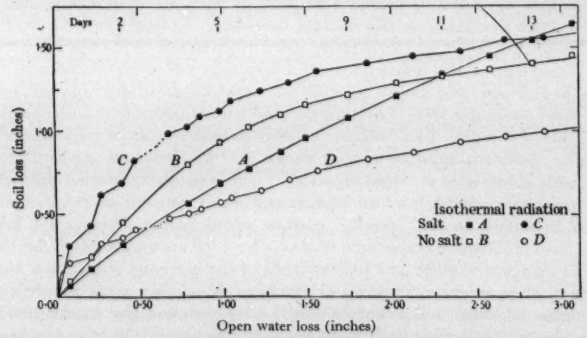


Figure 3. Effect of isothermal conditions, intermittent radiation and dissolved salt on evaporation from fallow soil.

agreement for winter periods was obtained showing that when the surface was moist, and upward flow of liquid could keep pace with the sink strength, the

order of magnitude of the evaporation rate was predictable from weather data. The limitations of the upward liquid movement were examined in laboratory experiments (Penman, 1941) in which radiation conditions characteristic of sunny June weather were imposed and the evaporation rates from irradiated surfaces and surfaces kept near air temperature were compared with open water losses. The results (figure 3) were very little different for a clay soil and for the sandy soil to which the figure refers. Strong irradiation of moist bare soil causes rapid initial drying with which upward movement through the soil cannot keep pace, and a thin dry surface layer is rapidly formed which imposes a big resistance to further drying. This " self-mulching " of a bare soil occurs under English summer conditions and is not limited to arid regions as suggested by Buckingham (1907). Fortier and Becket (1912) have provided similar data for a number of American soils. Decrease of the initial drying rate by reduced surface temperature (isothermal conditions) or by reduced surface vapour pressure (salt present) enables a much bigger deficit to be developed over an extended period ; it is under these conditions that artificial mulching by cultivation may reduce evaporation losses, i.e. in spring.

Statistical analysis of the Rothamsted records (Koshal, 1934) has given a set of 12 equations for the monthly evaporation from bare soil, of which the following are examples :—

$$\left.\begin{aligned} \text{Jan.}\quad & R-D_{20}=0{\cdot}398+(\mathit{0{\cdot}033}\pm 0{\cdot}038)R'-(\mathit{0{\cdot}0023}\pm 0{\cdot}0129)T',\\ \text{May}\quad & R-D_{20}=1{\cdot}578+(0{\cdot}389\pm 0{\cdot}041)R'-(0{\cdot}0633\pm 0{\cdot}0228)T',\\ \text{July}\quad & R-D_{20}=1{\cdot}965+(0{\cdot}504\pm 0{\cdot}040)R'-(\mathit{0{\cdot}0071}\pm 0{\cdot}0244)T'. \end{aligned}\right\} \quad \ldots\ldots (46)$$

In these equations $R-D_{20}$ is the difference between the rainfall and drainage (20 in. gauge) for the month in inches, R' and T' are the deviations from the long term mean monthly rainfall (in.) and mean monthly air temperature (°F.) respectively, and non-significant regression coefficients are given in italics. All regressions on rainfall from April to October are significant; the remainder are non-significant. Nearly all the regressi ns on temperature are non-significant, May being an exception, yielding the surprising result that the higher the May air temperature the less the May evaporation would be. A possible explanation is to be found in figure 3. A warm May probably means a sunny May, with evaporation following the " radiation " curve; a cool May probably means a cloudy May, with evaporation following the " isothermal " curve, leading to a bigger total evaporation. The constants in the above equations are the long period monthly mean values of $R-D_{20}$. As they show a seasonal variation with a January minimum and a July maximum, moving just a little out of phase with the solar radiation cycle, some correlation with other cyclic meteorological elements is to be expected, and if one chooses mean air temperature (Crowther, 1930) the result is

$$E=R-D=0{\cdot}07T-2{\cdot}22 \quad \ldots\ldots (47)$$

for any calendar month, where T(°F.) is the mean air temperature of the month. This equation does *not* mean that air temperature is an important element controlling evaporation; it merely states that evaporation and the warming of the air are two effects of the same cause (cf. (36)).

The limited depth to which evaporation effects can penetrate has been demonstrated by Keen (1931). Cylinders about 6 ft. deep and 2 ft. 6 in. diameter, with closed bottoms, were filled with soil of different types (clay, fine sand, coarse sand), sunk to lip level in an open site and left for some years to settle. In each a narrow gauge pipe was inserted and the daily movement of water level noted. After being water-logged in winter, spring evaporation exceeded rainfall so that the water-level sank, and figure 4 shows the progressive fall over

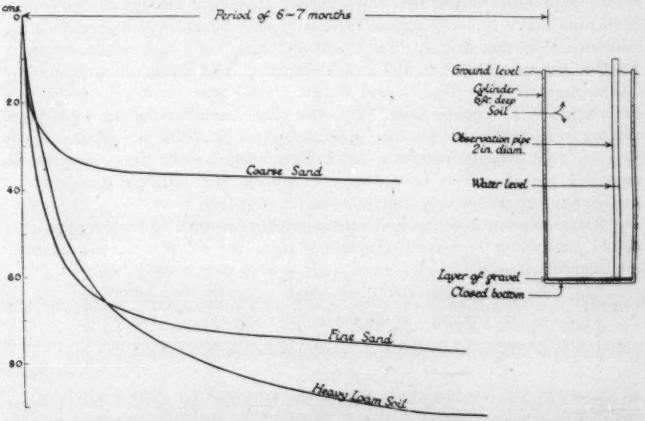


Figure 4. Retreat of water-table in soil, initially saturated, when continuous evaporative conditions are maintained at the surface.

a six-months' drought period in 1921. For the coarse sand, the effect of surface evaporation failed to reach more than 14 inches depth; for the clay soil, there was very little movement of water upward after the water-table had reached 3 feet. Similar results have been reported by Ridgeway (1902) and, for coarse river-bed materials, by Sleight (1917).

(iii) *Cropped soil* (*homogeneous areas*)

Transpiration from crops has been more a biological than a physical study, and a reasonable review of the physical aspects can only be made at the expense of some broad generalizations about the behaviour of plants. The principal new factors introduced are the upward growth of the aerial part of the plant from which the transpiration stream reaches the atmosphere, and the downward growth of the roots, enabling the plant to tap the water stored in a considerable depth of soil. Transpiration experiments in pots are unreliable guides to behaviour in the field because the sample plot is rarely typical of its environment, and this section will be confined to a review of results for extended areas of crop, or small areas of short crop such as grass. Thornthwaite and Holzman (1940) have applied their eddy diffusion formula (19), to estimating evaporation rates from grassland. Measurements at 2 ft. and 29 ft. above ground showed very small

differences in vapour pressure, and several hygrographs were needed at each level to obtain the necessary accuracy. Their daily records show the change over between the upward evaporation stream by day and the night down-stream leading to dew formation, but no direct observations of transpiration were made to set alongside the estimates.

The effect of the control of stomatal opening by daylight appears in results of White (1932) together with other points of interest. He worked in the long, narrow Escalante Valley, Utah, under semi-arid conditions where crops grew on only a small fraction of the total catchment area. Crops of lucerne (alfalfa), a very deep-rooting leguminous plant, could grow without irrigation on small areas in the valley bottom where the water-table (i.e. well level) was close to the surface. Well level showed daily fluctuations, the day fall being attributed to transpiration, and the night recovery attributed to slow seepage into the area of water from under bare areas not suffering the day transpiration loss. A typical record appears in figure 5, and shows a crop to be withdrawing water at about

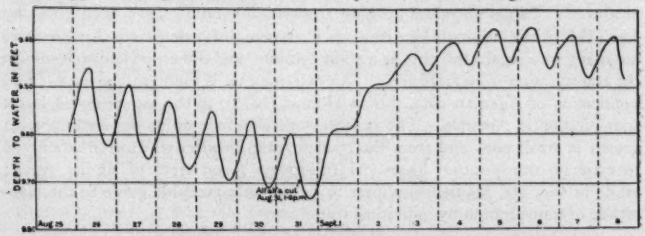


Figure 5. Daily movements of the water-table under a field of alfalfa (Lucerne).

10 ft. depth (cf. figure 4); shows the daily cycle and the steady daily movement until the crop was cut; shows the effective cessation of transpiration when the evaporating surfaces were removed; and shows its gradual resumption as the new growth began. White's calculations of transpiration rates were based on the assumption that the straight part of the rising night curve gave a measure of the natural rate of rise of the water-table in the absence of transpiration.

Over an area such as this there will be a " vapour blanket ", the pressure gradient in which Thornthwaite and Holzman attempted to measure. Bigelow (1908) found the height of this " blanket " to be about 30 ft., the same over a lucerne field as over a reservoir. It is apparent from this and other work that where a crop has an adequate water supply, the day transpiration rate is of the same order as the evaporation rate from an open water surface, and the plant appears to behave as a highly conducting channel between a source (soil water) and a sink (atmosphere). The source is extended in depth and its reserve of water will amount to the difference between the moisture content at field capacity and the moisture content at which the suction exerted by the roots can no longer overcome the negative pressure behind the highly curved menisci of the water held in the fine pores not yet emptied by root action. Schofield (1935) and co-workers have investigated the thermodynamics of this soil-water-plant equilibrium and shown that, measured in terms of the free energy of the soil

water, this limiting moisture content is, very closely, a true soil moisture constant, largely independent of soil or crop. When the reservoir of water upon which the root layer can draw is renewed sufficiently frequently by rain or irrigation, transpiration is not limited by moisture supply and summer behaviour of cropped sites is much the same as the winter behaviour of bare soil (figure 2). Data obtained by Lawes and Gilbert (1871) and Hendrick *et al.* (1921, 1938) indicate that transpiration losses are then dictated almost entirely by weather conditions and are largely independent of crop yield and fertilizer treatment.

The effect of type and depth of soil on transpiration appears in a collection of data for turfed drain-gauges and catchment areas in south-east England (Lewis, 1943). Soil type is comparatively unimportant, but deepening the soil from 24 in. to 48 in. increases the total annual transpiration by extending the reservoir available for root action.

The importance of water in crop growth has led many experimenters to express results in terms of the ratio of water transpired per unit weight of dry matter produced. Tables of values of this "transpiration ratio" have been given by Lee (1942), but although he dismisses it as being of little practical value there are many who regard the ratio as a plant constant, and use crop yields to estimate the transpiration from large areas. An extreme case is found in Horton's (1923) application of Austrian data, due to Höhnel (1879), to the estimation of forest transpiration in America. The latter's work was done on young seedling trees grown in small pots, and from the transpiration-growth data Horton attempted to estimate transpiration losses for full-grown forest trees of similar types; crude as they are, his figures appear to be the only available guide to the magnitude of transpiration by individual trees.

(iv) *Cropped soil* (*catchment areas*)

Estimates of evaporation from rainfall and run-off data are dependent on reasonable accuracy in gauging of these quantities, and on there being no deep percolation that cannot be measured. Topographical factors and type of plant cover are equally important with weather in determining transpiration. In America, Merriman (1907) has expressed monthly and annual evaporation in terms of mean air temperature and rainfall (cf. (46)), Kittredge (1938) has given a map of rainfall *minus* run-off for the United States and thereby provided the best available estimates for general forest transpiration, and Lee (1942) has given a diagram showing that when the annual rainfall exceeds 24 inches the annual evaporation loss from a number of American catchment areas is very nearly constant (cf. figure 2 : winter data).

In Great Britain, Lloyd (1936–42) has obtained statistical correlations between evaporation and mean air temperature, rainfall and sunshine, and has attempted to allow for deep percolation by including a geological term ranging from 8 in./year for cretaceous to zero for pre-cambrian and other impermeable rocks. For the seasonal variation of evaporation, Lloyd's expressions become somewhat complicated, and it is probable that the temperature term is merely expressing a seasonal change in the same way as it does in (47) above.

For large geographical areas of the earth, Murray (1887) gave annual values of evaporation loss which are generally about two-thirds of the corresponding

entries in table 1. Finally, for the largest catchments of all, Wüst (1922) has tabled three sets of estimates of the hydrology of the earth in which the main discrepancies occur in the data for oceanic rainfall and evaporation. The latest of the three indicates that the mean annual rainfall and evaporation over the whole earth are equal at 74·3 cm/annum, over the oceans the values are 73·9 and 84·2 cm/annum respectively, and over the land surface the values are 75·2 and 50·3 cm/annum respectively.

REFERENCES

Angot, A., 1883. *Ann. Bur. Cent. Met.*, Part I, 136.

Ångström, A., 1920. *Geog. Ann.*, **2**, 253.

Bagnold, R. A., 1941. *Physics of Blown Sand and Desert Dunes* (Methuen).

Banerji, S. K. and Wadia, H. M., 1932. *Mem. Ind. Met. Dept.*, **25**, 291.

Best, A. C., 1935. *M.O. Geophys. Mem.*, No. 65.

Bigelow, F. H., 1908. *Mon. Weath. Rev., Wash.*, **36**, 24.

Bilham, E. G., 1938. *Climate of British Isles* (Macmillan).

Bowen, I. S., 1926. *Phys. Rev.*, **27**, 779.

Briggs, L. J. and Shantz, H. L., 1913. *U.S. Plant Ind. Bull.*, **284**, 285.

Briggs, L. J. and Shantz, H. L., 1914. *J. Agric. Res.*, **3**, 1; 1915. *Ibid.*, **5**, 117; 1916. *Ibid.*, **5**, 583; **7**, 155; 1917. *Ibid.*, **9**, 277.

Brown, H. T. and Escombe, F., 1900. *Phil. Trans.*, **193**, 283.

Brunt, D., 1934, 1939. *Physical and Dynamical Meteorology* (Cambridge: University Press).

Buckingham, E., 1907. *Bull. U.S. Div. Soils*, No. 38.

Carpenter, L. G., 1889. *2nd Ann. Rep. Colo. Exp. Sta., p.* 49.

Carrier, W. H., 1921. *Ind. and Eng. Chem.*, **13**, 432.

Crowther, E. M., 1930. *Proc. Roy. Soc.* B, **107**, 1.

Cummings, N. W., 1940. *Trans. Amer. Geophys. U.*, 512.

Cummings, N. W. and Richardson, B., 1927. *Phys. Rev.*, **30**, 527.

Dalton, J., 1801. *J. Nat. Phil. Chem.*, **5**, 241.

Dalton, J., 1802. *Mem. Lit. Phil. Soc.*, **5**, 346, 535.

Dalton, J., 1834. *Meteorological Observations and Essays* (2nd edition).

Daubrée, G. A., 1847. *C.R. Acad. Sci., Paris*, **24**, 534.

Davydov, V. K., 1936. *Assoc. Inter. D'Hydrologie Sci.*, Bull. No. 22, p. 142.

Ebermayer, E., 1879. *Forstwiss. Centrbl.*, **23**, 77.

Eser, C., 1884. *Forsch. Geb. Agr. Phys.*, **7**, 1.

Ewald, P. P., Poschl, T. and Prandtl, L., 1930. *The Physics of Solids and Fluids* (Blackie).

Fitzgerald, D., 1886. *Trans. Amer. Soc. Civil Eng.*, **15**, 581.

Follansbee, R., 1934. *Trans. Amer. Soc. Civil Eng.*, **99**, 704.

Folse, J. A., 1929. *Carnegie Inst. Wash. Pub.*, 400.

Fortier, S. and Beckett, S. H., 1912. *U.S. Dept. Agric. O.E.S. Bull.* 248.

Freeman, J. R., 1926. *Regulation of Great Lakes.*

Frost, R., 1946. *Proc. Roy. Soc.*, A, **186**, 20.

Harding, S. T., 1922. *State Dept. Eng., California, Bull.* 9.

Hendrick, J., 1921. *Trans. Highland and Agric. Soc. Scot.*, **33**, 56; 1930. *Ibid.*, **42**, 1.

Hendrick, J. and Welsh, H. D., 1938. *Trans. Highland and Agric. Soc. Scot.*, **50**, 184.

Höhnel, F., 1879. *Forsch. Beg. Agrikultur Phys.*, **2.**

Holzman, B., 1941. *Trans. Amer. Geoph. U.*, 655.

Horton, R. E., 1923. *Mon. Weather Rev., Wash.*, **51**, 571.

Howard, L., 1837. *Seven Lectures on Meteorology* (Pontefract).

Jeffreys, H., 1918. *Phil. Mag.*, **35**, 273.

Katchinsky, N. A., 1934. *Trans. Inter. Soc. Soil Sci., Moscow*, **2**, 79.

Keen, B. A., 1931. *Physical Properties of Soil* (Longmans).

King, F. H., 1890–1. *8th Ann. Rep. Agric. Exp. Sta., Wisconsin.*

Kittredge, J., 1938. *J. Forestry*, **36**, 775.

Kohnke, H., Dreibelbis, F. R. and Davidson, J. M., 1940. *U.S. Dept. Agric. Misc. Pub.*, No. 372.

KOSHAL, R. S., 1934. *J. Agric. Sci.*, **24,** 105.
LAWES, J. B. and GILBERT, J. H., 1871. *J. Roy. Agric. Soc.*, **7,** 91.
LAWES, J. B., GILBERT, J. H. and WARINGTON, R., 1881–2. *J. Roy. Agric. Soc.*, **17,** 241, 311; **18,** 1.
LEATHER, J. W., 1913. *Mem. Dept. Agr. India (Chem. Series)*, **3,** 1.
LEE, C. H., 1942. *Physics of Earth—IX. Hydrology*, p. 259 (New York: McGraw-Hill).
LEWIS, W. V., 1943. *Proc. Geol. Ass.*, **54,** 171.
LIPPINCOTT, J. B., 1927. *Trans. Amer. Soc. Civil Eng.*, **90,** 356.
LIVINGSTON, G. L., 1908-9. *Mon. Weather Rev., Wash.*, **36, 37.** (In 8 parts, later issued as bound pamphlet.)
LLOYD, D., 1936. *Quart. J. Roy. Met. Soc.*, **62,** 219; 1940. *Ibid.*, **66** 181; 1941. *Ibid.*, **67,** 33; 1942. *Ibid.*, **68,** 35.
LUGEON, J., 1928. *Précipitations Atmosphériques* (Paris).
MASKELL, E. J., 1928. *Proc. Roy. Soc.*, B, **102,** 488.
MERRIMAN, T., 1907. *Mon. Weather Rev., Wash.*, **35,** 109.
MEYER, A. F., 1915. *Trans. Amer. Soc. Civil Eng.*, **79,** 1056.
MURRAY, J., 1887. *Scot. Geog. Mag.*, **3,** 65.
PASQUILL, F., 1943. *Proc. Roy. Soc.*, A, **182,** 75.
PENMAN, H. L., 1940. *Quart. J. Roy. Met. Soc.*, **66,** 401.
PENMAN, H. L., 1941. *J. Agric. Sci.*, **31,** 456.
PENMAN, H. L., 1943. *Quart. J. Roy. Met. Soc.*, **69,** 1.
PENMAN, H. L. and SCHOFIELD, R. K., 1941. *J. Agric. Sci.*, **31,** 74.
PEREIRA, H. C., 1941. *Emp. J. Exp. Agric.*, **9,** 29.
POWELL, R. W., 1942. *Trans. Instn. Chem. Engrs.*, **18,** 136.
POWELL, R. W. and GRIFFITHS, E., 1935. *Trans. Instn. Chem. Engrs.*, **13,** 175.
PRESCOTT, J. A., 1940. *Trans. Roy. Soc. S. Australia*, **64,** 114.
RAGONA, D., 1867. *Met. Ital. Sup.*, 13.
RAMAN, P. K. and SATAKOPAN, V., 1935. *Sc. Notes, India Met. Dept.*, **6.**
RIDGEWAY, C. B., 1902. *Wyoming Exp. Sta., Bull. No.* 52.
ROHWER, C., 1931. *U.S. Dept. Agr., Tech. Bull.*, 271.
ROHWER, C., 1934. *Trans. Amer. Soc. Civil Eng.*, **99,** 673.
SAHNI, P. N., 1941. *J. Agric. Sci.*, **31,** 110.
SCHOFIELD, R. K., 1935. *Trans. Third Inter. Congr. Soil Sci.*, **2,** 37.
SHAW, W. N., 1877. *Quart. Weath. Rpt. Meteorological Office*, 35.
SHAW, W. N., 1881. *Met. Council Rpt.*, 28.
SLEIGHT, R. B., 1917. *J. Agric. Res.*, **10,** 209.
STEFAN, J., 1881. *Sitzber. K. Akad. Wiss. Wien*, **83,** 943.
SUTTON, O. G., 1934. *Proc. Roy. Soc.*, A, **146,** 701.
SVERDRUP, H. V., 1945. *Oceanography for Meteorologists* (London: Allen and Unwin).
SYMONS, G. J., 1867. *British Rainfall*, 9.
TAYLOR, G. I., 1922. *Proc. Lond. Math. Soc.*, **20,** 196.
THORNTHWAITE, C. W. and HOLZMAN, B., 1939. *Mon. Weather Rev., Wash.*, **67,** 4.
THORNTHWAITE, C. W. and HOLZMAN, B., 1940. *Trans. Amer. Geoph. U.*, 510.
VISENTINI, M., 1936. *Assoc. Inter. D'Hydrologie Sci.*, Bull. No. 22, p. 119.
WHITE, W. N., 1932. *U.S. Dept. Interior.* Water Supply Paper 659.
WOLLNY, E., 1888. *Forsch. Geb. Agrik.-Physik*, **11,** 1.
WÜST, G., 1922. *Zeit. Ges. Erdekunde, Berlin.*

METEORS, COMETS AND METEORIC IONIZATION

By A. C. B. LOVELL, J. P. M. PRENTICE, J. G. PORTER, R. W. B. PEARSE, N. HERLOFSON

CONTENTS

	Page
Meteors, by J. P. M. Prentice	389
Comets and Meteor-Streams, by J. G. Porter	402
Spectroscopic Observations on Comets, Meteors and Meteorites, by R. W. B. Pearse	410
Meteoric Ionization and Ionospheric Abnormalities, by A. C. B. Lovell	415
The Theory of Meteor Ionization, by N. Herlofson	444

FOREWORD. The belief that the ionization due to meteors influenced the conditions in the E region of the ionosphere has been current since the early days of the ionosphere work. During the last few years, however, physicists have been able to deploy the refined radio techniques developed during the war, and the true significance of these effects is now appreciated. In addition to this it is now possible to study the effects of individual meteors by day and night under any weather conditions. Hence the integration of another branch of astronomy with physics has become essential ; the astronomer has to assimilate the new data on meteor activity in his study of the celestial mechanics of meteors and comets, and the physicist the existing knowledge of meteors gained by visual observations in order that he can begin to interpret the complex phenomena which are now revealed.

One of the first attempts to begin this integration was made at a Physical Society Conference held in the University of Manchester Physical Laboratories on 19 March 1947, and four of these Reports are based on lectures given there.* The Report by Dr. Pearse has been added in order to cover ground which was omitted at this Conference.

METEORS

By J. P. M. PRENTICE

§1. INTRODUCTION

THIS Report contains a brief introduction to the study of meteors, which have recently become of interest to the physicist because they can now be successfully observed by radio echo techniques. No subject can be compressed without omissions and distortions ; it is hoped that the distortions are only slight, but it has not been possible to do justice to the work, theoretical or observational, of men such as H. A. Newton, G. V. Schiaparelli and W. F. Denning, who laid the foundations of our knowledge of meteors. It is not our purpose here to consider the physical theories of meteors, which may be expected to undergo radical revision as a result of new knowledge, but to present the salient results of astronomical observation. It is a field in which results have been difficult to obtain, partly because all astronomical work is conditioned by clear weather and partly because of the inherent difficulty of the subject. Meteors are normally of such short duration and feeble luminosity and the observable

* An account of this Conference was given in *Nature, Lond.*, **160**, 76 (1947).

hourly number is generally so low that it is difficult to apply the normal techniques of astronomy, and in this work both the telescope and the photographic plate have hitherto been inferior to the unaided eye, considered purely as an optical system.

The radio echo technique will greatly increase our knowledge, for example in the physics of meteor phenomena, in the determination of meteor velocities and in the study of those meteor currents observable only in daylight; but it would be a mistake to think that the older techniques have become obsolete on this account. Each has its severe limitations and all are necessary in a difficult field of observation.

Visual Observations

The value of the visual technique of observation lies in its ability to determine with sufficient accuracy the positions of large numbers of meteors varying greatly in magnitude and angular velocity. Meteor streams have little cohesion and their individual meteors differ so greatly in their orbits (§ 3 (iii)) that conclusions cannot safely be drawn from merely a few observations, however refined. The methods of observation mainly used are those of the American Meteor Society (Olivier 1925, p. 9) and of the English observers (Prentice 1945) respectively, and these may be considered typical of modern practice. Usually this work has been carried out by observers working in isolation, but in those instances in which the same meteor is observed from two or more suitable places results of an altogether different order are obtained ; for example, certainty instead of inference in the derivation of radiants (§ 3 (iv)) and the determination of meteor heights and velocities (§ 5 (i)). The principal limitation of the visual technique is that the observer is unable to distinguish small intervals of time and hence to determine accurate velocities and variations in velocity. At present, too, the scale of visual observing is inadequate and expansion is necessary, particularly in the southern hemisphere where, despite the work of McIntosh (1929, 1934, 1936, 1940), large scale systematic meteor observations have hitherto been lacking.

§ 2. THE NUMBERS OF METEORS

(i) *The Major Showers*

The curve of variation in the hourly numbers of meteors throughout the year is dominated by strongly marked peaks of activity distributed at irregular intervals due to the incidence of certain meteor streams, the major showers. The following is a list (British Astronomical Association Handbook) of those annually visible in geographical latitude $+52°$.

Date of maximum	Name	Radiant R.A.	Radiant Dec.	Normal hourly rate
Jan. 3	Quadrantids	230°	+52°	30 to 40
Apr. 21	Lyrids	270	+33	7 ,, 10
Aug. 10–13	Perseids	47	+58	40 ,, 60
Oct. 20–23	Orionids	96	+15	10 ,, 20
Nov. 16–17	Leonids	152	+22	10 ,, 15
Dec. 11–13	Geminids	113	+32	60

Outside the period of visibility of the major showers there is a general level of activity of about 5 to 10 meteors an hour, and this is maintained in part by a

number of very feeble streams (the " minor streams ") and in part by single meteors which, as a result of perturbations, are pursuing an isolated path in space. The minor streams are believed to be very numerous, but owing to their tenuity (the richest rarely produce more than one meteor in an hour) they are difficult to study. Certain of the major showers have occasionally produced " meteor storms " when meteors have been described as falling like snowflakes ; records of these are found in the old chronicles, particularly of the ancient Chinese (Biot 1841, Olivier 1925), which carry meteor records back to several hundred years B.C. These storms are of very rare occurrence ; the only instances during the past 120 years have been the great Leonid showers of November 1833 and 1866, the Bielid showers of November 1872 and 1885, and the Giacobinid showers of October 1933 and 1946.

(ii) *Radiant Altitude*

The *observed* hourly rate of a given shower will not remain constant for constant intensity, but will vary with the altitude of the radiant, i.e. of the point on the celestial sphere from which the meteors are directed (see § 3). The law of variation cannot be rigorously given and will vary with the distribution of magnitude (luminosity) of the meteors in the stream, but an approximate expression is

$$R_0 = R \sin (a+6)^\circ,$$

where R is the true rate, R_0 is the observed rate and a is the radiant altitude. For $a = 6^\circ$, 11°, 18°, $\sin(a+6)^\circ$ has values of 0·2, 0·3, 0·4 respectively; when the radiant is at a low altitude it is therefore impossible to obtain good results even at maximum.

(iii) *The Minor Streams*

If we eliminate the major showers and study the curve of variation of the general level of meteoric activity we find firstly a daily and secondly an annual variation (Olivier 1925, pp. 55, 182) as follows : (1) The hourly rate increases throughout the night, with a minimum at 00^h *minus* 6 hours Local Mean Time and a maximum at 00^h *plus* 6 hours or nearly so. (2) The hourly rate at corresponding L.M.T. varies throughout the year, with a minimum in the spring (in European latitudes) and a maximum in the autumn. Both sets of variation are due to the same cause, of which the following is an idealized explanation. If we assume that meteor radiants are uniformly distributed over the celestial sphere and are of equal intensity, then, if the Earth were at rest, the hourly number visible at a given point would be constant. But the Earth moves in its orbit around the Sun with a mean velocity of about 30 km/sec., and hence by the combination of velocities the apparent radiants will be concentrated about the point on the celestial sphere to which at a given time the Earth's motion is directed along the tangent to its orbit. This point, known as the " Apex of the Earth's way " or " Meteoric Apex ", is a point on the ecliptic approximately 90° from the Sun and its declination therefore varies in the course of the year from $-23^\circ{\cdot}4$ to $+23^\circ{\cdot}4$. Now we have seen that the *observed* hourly rate of a major shower is a function of the altitude of its radiant; similarly, since the radiants are concentrated towards the Apex the general level of meteor activity is a function

of the altitude of the Apex. The annual variation in the altitude of the Apex at equivalent times, due to its varying declination, may be illustrated by the values attained in Spring (15°) and in Autumn (61°) when the Apex is on the meridian in geographical latitude +52°. The *daily* variation in the altitude of the Apex in the same latitude and at mean declination is between −38° in the evening and +38° in the morning, a daily range of 76°. These variations in the altitude of the Apex account in broad outline for the variation in the daily and seasonal numbers of the mean level of meteor activity. The agreement is never exact, and this is due to the fact that our basic assumption, the uniform distribution and richness of meteor radiants, is never fulfilled in practice.

(iv) *The True Numbers of Meteors*

We have assumed hitherto that a meteor observer sees all the meteors which appear within his field of observation, but owing to the imperfections of the human eye the facts are far otherwise. The observer can detect a bright meteor at a considerable angular distance from the centre of his field of vision, but with decreasing luminosity the diameter of the effective field becomes very restricted. For example, Backhouse (1884) gives the following approximate figures :

Magnitude	−4·5	+1	+3	+5
Diameter of effective field	100°	65°	45°	25°

Similar results have been obtained by Öpik (1923) and other investigators in the course of statistical studies with groups of observers watching a common field. Thus, although the *observed* distribution of meteor magnitudes is as shown by the characteristic figures given in line (1) of table 1, the general feature of which is a steady increase in numbers with decreasing luminosity down to magnitude +4 with a rapid decline thereafter, the real numbers which an observer could see given optical perfection in the eye are those shown in line (2).

Table 1

Magnitude	⩾1	2	3	4	5	6	Σ
(1) Observed number	5	9	20	30	26	10	100
(2) True number	5	11	25	58	133	306	538

Similar work on meteors too faint to be visible with the unaided eye has yielded similar results for telescopic meteors, the peak of the magnitude curve depending on the light grasp of the instrument employed. It has been shown furthermore that the gradient of the ascending curve is the same in both classes (Williams 1939), i.e. the rate of increase in numbers is constant over a range of at least 6 or 8 magnitudes. We may therefore form the following rough estimate of the true numbers of meteors on an ordinary non-shower night (*a*) in the field of observation of one observer and (*b*) over the whole Earth (Watson 1941) (table 2).

Table 2

Limiting magnitude	+1	+6	+10
True hourly rate in field of one observer	0·5	45	1875
True daily rate in field of one observer	11	1100	45000
True daily rate over whole Earth	$1{\cdot}1\times10^8$	$1{\cdot}1\times10^8$	45×10^8

§3. METEOR RADIANTS

(i) *Significance*

During the great Leonid shower of 1833 Olmsted (1834) and Twining (see Olivier 1925, p. 24) noticed that the meteors were emanating from a point in the constellation Leo which was fixed relative to the stars. This they correctly explained as showing that the meteors were moving in parallel paths in the Earth's atmosphere, the apparent divergence from a point being an effect of perspective, and they therefore concluded that the meteors were members of a single stream having its origin outside the Earth's atmosphere. But the observed point of radiation of a meteor stream is only the apparent position since the motion of a meteor in the Earth's atmosphere is along the resultant of two velocities, due to the orbital motions of the meteor and of the Earth respectively. The true radiant will therefore lie further from the Apex of the Earth's Way, the angular displacement $\epsilon' - \epsilon$ being a function of the relative velocity and of the angle between the tangents to the two orbits at the point of intersection. If we assume parabolic velocity the angle $\epsilon' - \epsilon$ has the following values :

ϵ	0°	30°	60°	90°	120°	150°	180°
$\epsilon' - \epsilon$	0	21	38	45	38	21	0

Here the angle ϵ is the " elongation " of the apparent radiant, i.e. the angular distance of the apparent radiant from the Apex of the Earth's Way. Thus if we either know or can assume the velocity of the meteors we know the orientation in space of the tangent to the meteor orbit at the point where it intersects that of the Earth, and hence we can derive the elements of the meteor orbit : inclination, perihelion distance, eccentricity etc. In this way the connection between meteors and certain comets was finally established in the years 1866–67 (Clerke 1893, Olivier 1925, p. 29).

(ii) *Moving and Stationary Radiants*

Since the position of the apparent radiant is conditioned by two vectors, of which that due to the orbital motion of the Earth varies from day to day by approximately 1° a day in increasing longitude, we should expect to find a daily displacement in the observed radiant of any stream which is visible over a period of several days. The classic illustration of this is the great Perseid shower (Denning 1899), whose mean apparent radiant (assumed to be concentrated in a point) moves from 28° + 53° on July 28 to 53° + 59° on August 17. The daily motion of the radiants of other streams has also been demonstrated. The amount of the daily motion will not be uniform over the celestial sphere, but will be a function of the celestial longitude λ and latitude β (von Niessl 1880, Olivier 1925, p. 101) ; for parabolic velocity a special case of peculiar interest occurs when $\lambda = \odot + 144\frac{3}{4}°$, $\beta = 0°$, where $\odot$ is the longitude of the Sun, for here the motion of the radiant is approximately zero over a whole month. Several minor streams are known for which this condition is approximately fulfilled, and because they are of long duration they exercise a disproportionate influence on the general distribution of meteor radiants.

(iii) *Types of Radiant*

The conception that the meteors of a given stream move in parallel paths is, however, only an ideal model which is commonly found in general text books but rarely in actual observation. The Leonids certainly have a sharp radiant, but those of all the other annual major showers are diffused over a considerable area ; for example the radiants of the individual meteors of the Perseid stream are scattered over an ellipse approximately 20° in major axis and 15° in minor axis. We recognize several types of radiant structure, for example (*a*) a point radiant of the classic type, as the Leonids; (*b*) radiation from a considerable area, but mainly concentrated in sharp sub-centres, as the Orionids (Prentice 1933–39); (*c*) radiation diffused fairly uniformly over a considerable area, as the Perseids. The existence of these large areas of meteor radiation necessarily implies that there is a very great want of parallelism between the individual meteors of a given stream, and yet it is clear from their simultaneous appearance, their high rate of occurrence and the restricted area of radiation, that they are members of the same stream, presumably having a common origin. Hence, most streams clearly do not consist of particles moving in identical or even closely similar orbits, so that at distances greatly exceeding that of the Earth from the Sun they would cease to be recognizable as a stream at all. We might at first consider that these diffuse radiants are due to the secular disintegration of a unified stream through planetary perturbations, but it seems questionable whether perturbations can account for all the observed features of diffuse radiants.

(iv) *Determination of Radiants*

There are two methods by which radiants may be determined, firstly from the observed divergence of a number of meteors from a certain point on a certain date, by which we infer that they are members of one stream, and secondly from the parallactic displacement of the apparent paths of a meteor in those instances in which the same meteor is observed from two or more places as described in § 5 (i). In the latter method the radiant is known with certainty within the limits of observational error, but in the former it can only be inferred because the angle of foreshortening of each individual meteor cannot be directly determined. This introduces an uncertainty in the determination of group radiants which is not serious in the case of the major showers because of their dominant numbers, but becomes appreciable in the case of minor streams (Maltzev 1928, Öpik 1934, Prentice 1941). As most determinations of radiants in the past have been made by this method of group radiants, the alternative technique of detecting minor streams by duplicate observations of individual meteors is obviously of great importance.

§ 4. METEOR VELOCITIES

(i) *Significance*

For motion in an elliptic orbit the velocity at any point will depend upon the distance of the meteor from the Sun and on the semi-major axis a of the orbit. In a parabolic orbit $a = \infty$ and the parabolic velocity at 1 unit (the mean distance of the Earth from the Sun) is 42·1 km/sec. If the velocity at unit distance exceeds this figure the orbit is hyperbolic and the meteor cannot be, or cannot remain, a member of the solar system. We are therefore critically

concerned with the velocities of meteors, as any marked preponderance of hyperbolic velocities would show that these bodies are not members of the solar system, but come from interstellar space.

It is, however, very difficult to determine these velocities with the necessary accuracy, partly owing to the difficulties of technique and partly because a very small difference in the observed velocity will make a very large difference in the major axis of the orbit. Compare, for example,

Observed velocity	41·0 km/sec.	42·1 km/sec.
Major axis of orbit	18 units	∞
Period	27 years	∞

If, therefore, the majority of meteors are moving in elliptic orbits of great eccentricity (as we should expect by analogy with comets) having periods varying from, say, twenty to several hundred years, a difference of only 0·1 km/sec. in the observed velocity will materially alter the form of the orbit. No technique at present available can determine the velocities with such accuracy, and it is usual therefore to assume parabolic velocity as this greatly simplifies the calculation of the orbit ; it is understood that this is purely a convention adopted for convenience of calculation. Furthermore, it is erroneous to assume that the meteors of a given stream will all have the same velocity, for if we assume, as is currently held, that a meteor stream is caused by the gradual dissipation of a knot of cometary matter around an orbit, the very process of dissipation implies differing velocities. We have noted above (§ 3 (iii)) that a meteor stream is not a unified structure, but consists of particles pursuing independent orbits in space, so that the velocity will vary from one particle to another just as do the other orbital elements. Hence the most that we can expect to establish in determining the velocity of a meteor stream is a statistical mean of the velocities of the individual particles.

(ii) *Heliocentric and Geocentric Velocities*

The velocities referred to in the preceding paragraph are the heliocentric velocities, i.e. the velocity of the meteor in its orbit, and are to be distinguished from the geocentric velocities, i.e. the relative velocity of the meteor with reference to the moving Earth, and taking account of the acceleration due to the Earth's gravity. The unaccelerated heliocentric and geocentric velocities are related by the usual formula connecting the components of the parallelogram of velocities, and the increment due to the Earth's acceleration varies, for parabolic velocity, from 1 to 5 km/sec. according to the angle ϵ. Meteor velocities are by convention usually given as observed, viz. *geocentric* velocities, in the same way that radiants are usually given as observed, *apparent* radiants, but it is understood that in discussing the origin of meteors the observed velocities are to be converted into heliocentric.

(iii) *Determination of Velocities*

Before the development of radar technique there were three principal methods of determining meteor velocities, two by direct observation and one by indirect. The direct methods were :

(*a*) By determining the linear length of path and the duration of flight. The linear length of path P may be determined in those cases in which a meteor is observed from two places suitably spaced (§ 5 (i)), and the duration of flight t

is one of the data recorded by the observer ; we thus have $V = P/t$ as one of the elements of the true path derived from the duplicate observation. The duration t is obtained in visual work by estimation and in photographic work by the use of an occulting sector, and in both cases is capable of showing if the observed velocity differs greatly from parabolic. For example, Whipple (1940, 1947) has obtained a good photographic determination of the velocity of the November Taurids of approximately 30 km/sec. (geocentric), corresponding to a period of 3·3 years, and of 36 km/sec. for the Geminids, equivalent to a period of 1·7 years. These two streams have exceptionally short periods and neither technique gives sufficiently accurate data for meteors of normal velocity.

(*b*) The second direct method is by the observation of periodic recurrences of a rich stream, from which the mean period, and hence the velocity, are derived with considerable accuracy. This is in fact the only reliable method in ordinary cases, but its value is strictly limited by such factors as bad weather, the long period of most streams, in some cases the even distribution of the meteoric material around the orbit and, in the case of the minor streams, their extreme tenuity. It thus happens that very few velocities have been determined by this method ; the only well authenticated cases are given in table 3.

Table 3

Stream	Date	Principal Observed Recurrences	Period
Leonids	Nov. 16	1799*†, 1833*, 1866*, 1932	33·2 years
Bielids	Nov. 14‡	1838, 1872*, 1885*, 1892, 1899	6·6
Giacobinids	Oct. 9	1926, 1933*, 1946*	6·6

* Meteoric storm. † Radiant assumed.
‡ Date changes rapidly owing to regression of node.

(*c*) The indirect method of determining meteor velocities is as follows (Olivier 1925, p. 184). We have seen that the hourly number of meteors visible at a given place and time is a function of the altitude of the Apex and that this is due to the concentration of apparent meteor radiants around the Apex; furthermore, that this concentration arises from the combination of two velocities of which that of the meteors is one. Conversely, therefore, by observing the hourly numbers, we can infer the mean meteor velocity. If the mean velocity is markedly elliptical the concentration around the Apex will be greater and the gradient of the curve of hourly variation will be steeper, and the reverse if the mean velocity is markedly hyperbolic. Clearly it is only the *mean* meteor velocity which is derivable by this method, not the velocity of an individual particle or the mean velocity of an individual stream such as is yielded by the first and second direct methods respectively. This indirect method is that by which the supposed preponderance of hyperbolic velocities was originally inferred and is open to criticism which will be developed in considering the hyperbolic theory (§ 4(iv)).

(iv) *The Mean Velocity of the Non-shower Meteors*

Two views are currently held concerning the mean velocity of the non-shower meteors, some workers considering that they have enormously hyperbolic velocities and come from interstellar space, others considering that they are all, or nearly all, members of the solar system.

(*a*) The hyperbolic theory has its roots in the work of von Niessl (see Olivier 1925, p. 184) who sought to find a better accord between theory and observation in the variation in the hourly numbers of meteors. We have seen that if we assume a velocity in excess of parabolic the apparent radiants will not be so concentrated around the Apex, and by comparison with observation von Niessl found that better agreement was obtained if the mean velocity was taken as considerably greater than parabolic. This work was continued by Hoffmeister (1922, 1924, 1936) who carried out meteor counts in various latitudes and developed the theory of hourly variation so that he was able to evaluate the mean meteor velocity from the observed hourly rates ; this he determined at the enormously hyperbolic value of 90 km/sec. He also concluded from irregularities in the daily and annual variation that there exist preferential directions of motion in the hyperbolic meteors, and following in part the work of von Niessl and Knopf he inferred the existence of two great interstellar meteor currents, one directed from the constellation Scorpio and the other from Taurus. More direct observational support for the mean hyperbolic velocity of meteors was supplied by the results of the Harvard-Cornell Meteor Expedition in Arizona in 1931–33 (Shapley *et al.* 1932, Öpik 1934, 1937, Boothroyd 1934). A very extensive series of both visual and telescopic observations was obtained, and by making certain assumptions as to the mean centres of radiation the mean velocities of various groups of meteors could be determined. The instrumental equipment included an ingenious rocking mirror (Öpik 1934 b) which imparted a helical distortion of the locus of the meteor in flight, the duration of flight being determined from the form of the helix. The conclusions reached from this work were that meteors are of two general classes, shower and non-shower, of completely different characteristics ; that the non-shower meteors greatly outnumber the shower meteors ; that the non-shower meteors have greatly hyperbolic velocities, and that the velocity increases with decreasing luminosity. The markedly hyperbolic velocities obtained and their distribution according to luminosity are illustrated by the following figures (Öpik 1934) for the percentage of heliocentric velocities exceeding 62 km/sec., i.e. roughly 1·5 times parabolic :

Magnitude group	2·5	4·5	6·5	8·5
Percentage > 62 km/sec.	30	45	56	87

The peak of the heliocentric velocity curve for magnitude group 6·5 was found at approximately 100 km/sec. and that for magnitude group 8·5 was sharply marked at 143 km/sec. At the same time the heights in the atmosphere of these high velocity meteors were found to be of the same order as for shower meteors (Öpik 1937) (in fact, rather less), which caused some trouble since the dominant factor affecting meteor heights is the geocentric velocity (§ 5 (ii) below). The difficulty was resolved by the somewhat gratuitous assumption of a difference in chemical composition between the shower and non-shower meteors.

(*b*) The work on hyperbolic velocities has many attractive features but is open to serious criticism. That of von Niessl and Hoffmeister, on which it was originally based, depends largely on an indirect method whose fundamental assumption, the uniform distribution and richness of meteor radiants, is known to be invalid. It is also opposed to the results of many other observers based on

direct observations of position, and without the support of the Arizona work it could not therefore stand. There is also a direct conflict of observation between the Arizona results and those of other observers, American (Olivier 1938) and English, which we may illustrate by a recent analysis by Porter (1943, 1944) of British meteor data. A fuller account of this is given in § 5, but we may summarize his results as follows : (i) there is no important difference between shower and non-shower meteors, all of which have similar heights of appearance and disappearance in the atmosphere ; (ii) the observed velocities are less than parabolic and are similar in both classes and all meteors are therefore members of the solar system ; (iii) the observed velocity *decreases* with decreasing luminosity. Thus the observational disagreement could not be more complete. In this impasse the writer's opinion is as follows. Firstly, the work of the Arizona Expedition is open to criticism on several grounds, notably that they used a base line (approximately 36 km.) which was too short and that their technique of observation was not sufficiently accurate to enable them to determine individual radiants, which was the major purpose of their work. They therefore had to make mere assumptions as to the probable mean centres of radiation. Secondly, the uniformity of meteor heights for the shower and non-shower meteors (§ 5 (iii)) indicates that there is no considerable difference in the velocities of these two groups. For these and other reasons too detailed for this summary the writer considers the whole theory of general hyperbolic meteor velocities to be erroneous ; but this opinion would not be accepted by all meteor workers.

§ 5. PATHS IN THE ATMOSPHERE

(i) *Principles of Determination*

An observer working alone will only record the apparent path of the meteor as he sees it projected against the star background, i.e. he cannot determine its distance from him, and hence its height, nor can he determine the angle of foreshortening, and hence the radiant. But where the same meteor is seen from two places suitably situated, the parallactic displacement of the two apparent paths enables these to be determined. The data yielded by a duplicate observation include the radiant, the heights of beginning and end of luminous path (H_1, H_2), the linear length of path (P) and the observed velocity (S), together with the observational errors in all these quantities except the radiant. If the meteor is observed from three or more places we also determine both the error in the radiant and the personal errors of the observers. Various methods exist for computing the true paths of meteors (Schaeberle 1895, Olivier 1925, Fisher 1933, Whitney 1936, Davidson 1936, Porter 1943 a); of these the method of Davidson and Porter is well suited to the technique of observation and has been largely used in the systematic reduction of observations. The mean observational errors in the best modern visual work are of the order of 1° in radiant and 2–4 km. in height (Porter 1943 b). The most extensive modern work on the paths of meteors in the atmosphere has been that carried out by the Arizona Expedition and by the English meteor observers respectively (Denning 1897, 1912, 1916, Porter 1943–44), but the agreement between the two is not good, and in this discussion the results of the latter have been preferred for the reasons previously given.

(ii) *Meteor Heights*

The meteor zone of the atmosphere lies between heights 130 km. and 70 km. approximately, although occasional exceptional meteors exceed these limits. The heights at which meteors appear and disappear are determined mainly by (*a*) the initial geocentric velocity, which is the dominant factor (Porter 1944), and (*b*) the standard magnitude (M) (Öpik 1937), which is defined as the observed magnitude reduced to a standard distance of 100 km. in the zenith, thus eliminating the varying effects of distance and altitude.

(*a*) The effect of initial velocity V on the heights H_1, H_2 at the beginning and end of the path is well illustrated by the mean heights of the shower meteors which, owing to their different elongations, have different geocentric velocities. These are given in table 4; the mean standard magnitudes are also given in order to show that they are approximately constant in this sample.

Table 4

Shower	M	ϵ (deg.)	V km/sec.	H_1 km.	H_2 km.
Leonids	−1·8	10·5	72	125·6	94·1
Orionids	+1·2	24·1	68	119·7	100·9
Perseids	+0·5	39·2	61	114·9	94·7
δ Aquarids	+0·6	69·4	44	101·1	89·0
Taurids	−0·3	81·4	36	101·6	77·7
α Capricornids	+0·8	93·0	30	95·5	85·3

The same effect is shown in the grouped mean heights of the non-shower meteors, thus :

V (km/sec.)	70	61	45	32	20
H_1 (km.)	116	113	104	91	91
H_2 (km.)	101	93	85	76	70

(*b*) The influence of standard magnitude on height is much smaller and is shown by this analysis of the heights in the Perseid shower, the effect of V being thus largely eliminated.

M	−3·4	−1·1	+0·5	+1·8	+3·4
H_1 (km.)	116	115	111	112	109
H_2 (km.)	87	95	96	99	98

The bright meteors (negative values) clearly begin at a greater height and end at a lower height than the faint.

(iii) *Shower and Non-shower Meteors*

Since the initial velocity V exercises a dominant influence on the heights, a comparative study of the heights of shower and non-shower meteors has an important bearing on the problem of the supposed hyperbolic velocities. A simple comparison of shower and non-shower meteors cannot be made because it cannot be assumed that in a given sample V is distributed at random, and it is considered that the difference in height between the two groups announced by Öpik (1937) (in the sense shower > non-shower) is due to faulty sampling (Porter

1944). If we select equivalent groups of shower and non-shower meteors classified according to ϵ (and hence V) we have the comparison shown in table 5.

Table 5

ϵ	V km/sec.	Shower Meteors Shower	H_1 (km.)	H_2 (km.)	Equivalent Non-shower H_1 (km.)	H_2 (km.)
11°	72	Leonids	126	94	121	101
24	68	Orionids	120	101	114	101
39	61	Perseids	115	95	112	94
69	44	δ Aquarids	101	89	100	84
81	36	Taurids	102	78	99	78
93	30	α Capricornids	96	85	93	81

We conclude therefore that all meteors, whether belonging to the major showers or not, appear and disappear at equivalent heights in the atmosphere, and that there is no difference in the behaviour of the two groups.

(iv) *Retardation*

The observed velocities of meteors derived from P/t are consistently smaller than the parabolic velocity; this is illustrated in table 6, in which the mean observed velocities (S) of six shower and five typical non-shower groups are compared with the parabolic velocities (V).

Table 6

Name	ϵ	V (km/sec.)	S (km/sec.)	Ratio (S/V) %
Shower meteors				
Leonids	11°	72	58	81
Orionids	24	68	52	76
Perseids	39	61	43	70
δ Aquarids	69	44	29	66
Taurids	81	36	26	72
α Capricornids	93	30	24	80
Non-shower groups				
I	18	70	50	71
III	41	61	45	74
V	67	45	31	69
VII	90	32	21	66
IX	131	20	18	90

The observed velocities are some 25% below the parabolic; and this might be attributed to either (*a*) markedly elliptic velocity, (*b*) a retarding force, or (*c*) scale errors. The first of these is unlikely *a priori* and there are good reasons for rejecting the third (Porter 1944). The defect in velocity, which we therefore assume to be due to a retarding force, is correlated with the standard magnitude; thus if we take the ratio of the observed velocity S to the parabolic V as the most suitable basis of comparison, in order to eliminate the linear deceleration, we have

M	−4·2	+0·1	+1·2	+2·3	+3·9
S/V	0·92	0·75	0·66	0·62	0·53

The visual observations therefore clearly indicate a retarding force, but they can of course only yield mean values. By photographic techniques, however,

measurable retardations may be observed in individual meteors (table 7) (Millman and Hoffleit 1937).

Table 7

Meteor number	Number of break in trail								
	20	21	22	23	24	25	26	27	28
7	4·0	6·8	6·4	6·6	6·4				
14	3·3	3·0	3·4	3·6	4·5	4·8	5·4	5·4	8·8

The figures given are the deceleration in km/sec^2 at various points along the photographed trail. From similar studies of photographed shower meteors Millman and Hoffleit concluded that the Leonids and Perseids suffered a decrease in velocity of at least 10 km/sec. and the Orionids probably even more. These results are at variance with the radar work of Hey (Hey *et al.* 1947) on the Giacobinid shower in 1946, but upon this there are two points to be noted ; firstly, Hey's technique depends on the abnormal reflection of the radio waves and can only be used in a few instances presumably of very large and therefore very bright meteors (for which $S/V=1$, nearly), and secondly the Giacobinids have the smallest geocentric velocity of any of the major showers, and hence the linear decrease in velocity would be a minimum. This material therefore is the least suitable for the study of retardation and it may be hoped that with a less restricted radar sample the different results of the two techniques will be brought into agreement.

REFERENCES

Backhouse, T. W., 1884, *Observatory*, **7**, 299.

Biot, E., 1841, *Catalogue Générale des Étoiles Filantes observées en Chine pendant 24 Siècles.*

Boothroyd, S. L., 1934, *Harvard Coll. Obs. Circular*, No. 390.

British Astronomical Association Handbook, published annually.

Clerke, A. M., 1893, *History of Astronomy during the Nineteenth Century* (London : A. and C. Black), pp. 397, 403.

Davidson, M., 1936, *Brit. Astr. Assocn. J.*, **46**, 292.

Denning, W. F., 1897, *Mon. Not. R. Astr. Soc.*, **57**, 164 ; 1899, *Mem. R. Astr. Soc.*, **53**, 210 ; 1912, *Mon. Not. R. Astr. Soc.*, **72**, 424 ; 1916, *Ibid.*, **76**, 219.

Fisher, W. J., 1933, *Proc. Nat. Acad. Sci.*, **19**, 209.

Hey, J. S., *et al.*, 1947, *Nature, Lond.*, **160**, 74 ; *Mon. Not. R. Astr. Soc.*, **107**, 176.

Hoffmeister, C., 1922, *Astr. Abhandlungen*, **4**, No. 5 ; 1924, *Astr. Nachr.*, No. 5302–3 ; 1936, *Das interstellare System der KleinKörper* (Berlin).

McIntosh, R. A., 1929, *Trans. N.Z. Inst.*, **60**, 448 ; 1934, *Ibid.*, **63**, 433 ; 1936, *Trans. Roy. Soc. N. Z.*, **66**, 60 ; 1940, *Ibid.*, **69**, 392.

Maltzev, V. A., 1928, *Astr. Nachr.*, **234**, 243.

Millman, P. M., and Hoffleit, D., 1937, *Harvard Coll. Obs. Annals*, **105**, 611/2.

von Niessl, G., 1880, *Theoretische Untersuchungen über der Verschiebungen der Radiationspunkte aufgeloster Meteorstrome* ; 1917, *Smithsonian Miscellaneous Collections*, **66**, No. 16.

Olivier, C. P., 1925, *Meteors* (Baltimore : Williams and Wilkins) ; 1938, *Popular Astron.*, **46**, 325.

Olmsted, D., 1834, *Amer. J. Sci.*, **26**, 132.

Öpik, E., 1923, *Publ. de l'Observatoire de Tartu*, **25**, 1 ; 1934 a, *Harvard Coll. Obs. Circular*, No. 388 ; 1934 b, *Ibid.*, No. 389 ; 1934 c, *Ibid.*, No. 390 ; 1934 d, *Ibid.*, No. 391 ; 1937, *Harvard Coll. Obs. Annals*, **105**, pp. 549, 574, 588.

Porter, J. G., 1943 a, *Mem. Brit. Astr. Assocn.*, **34**, 37 ; 1943 b, *Mon. Not. R. Astr. Soc.*, **103**, 134 ; 1944, *Ibid.*, **104**, 257.

Prentice, J. P. M., 1933, *Brit. Astr. Assocn. J.*, **43**, 376 ; 1936, *Ibid.*, **46**, 329 ; 1939, *Ibid.*, **49**, 148 ; 1941, *Ibid.*, **51**, 164 ; 1945, *Brit. Astr. Assocn. Handbk.*, p. 36.

Schaeberle, J. M., 1895, *Cont. Lick Obs.*, No. 5.

SHAPLEY, H., *et al.*, 1932, *Proc. Nat. Acad. Sci.*, **18**, No. 1.
WATSON, FLETCHER, 1941, *Between the Planets* (Philadelphia : Blakiston Co.).
WHIPPLE, F. L., 1940, *Proc. Amer. Phil. Soc.*, **83**, 711 ; 1947, *Ibid.*, **91**, 189.
WHITNEY, B. S., 1936, *Mon. Not. R. Astr. Soc.*, **96**, 544.
WILLIAMS, J. D., 1939, *Astronom. J.*, No. 1110.

COMETS AND METEOR-STREAMS

BY J. G. PORTER

OUR knowledge of the relationship between certain comets and their associated meteor-streams may be said to date from the work of H. A. Newton on the Leonids in 1863. An excellent account of this early work and of further details of the subject up to 1930 will be found in two books by Olivier (1925, 1930). This Report is intended to summarize present knowledge of certain aspects of the subject, especially perturbations, which have a bearing on recent radio-echo measurements of meteor-streams.

§1. ORBITS

Since the average comet is seen for only a short period, it is often a matter of difficulty to determine an accurate orbit, for the small part of the comet's path that is seen forms only a short arc near perihelion, in which there is little apparent difference between the curvature of a parabola, hyperbola, or ellipse of large eccentricity. For this reason, and because the calculations involved are more simple, the orbits of many comets are given as parabolae, but it is understood that a comet cannot actually travel in a parabolic orbit, for the slightest perturbations (and these are always present) will transform such an orbit into an ellipse or hyperbola. Where prolonged observations are possible, a more accurate elliptical or hyperbolic orbit may usually be assigned to the comet.

The existence of comets with hyperbolic orbits would seem to imply that these bodies have an extra-solar origin, since the velocity in the orbit is greater than parabolic ; but there are definite reasons for rejecting this view. No orbits are known in which the hyperbolic character is very marked, the eccentricity in all cases being only very slightly greater than unity. Again, comets do not show any preferential direction of approach to the Sun. If they came from interstellar space, there would be a definite tendency for more comets to approach the Sun from the direction of the solar apex, but this is not the case. The most convincing argument comes, however, from an investigation of the past history of all comets having well-determined hyperbolic orbits. The motion of the comet is calculated backwards in time for a number of years, allowance being made for planetary perturbations, and in every case it has been found that the comet was originally travelling in an elliptical orbit, which has been altered by the perturbations due to the major planets (Strömgren 1914 ; a complete list of references is given by Merton 1947).

Thus we have every reason to believe that all comets are, like meteors, members of the solar system, travelling in elliptical orbits about the Sun. These orbits vary greatly in size, shape and direction in space. The periods range from the 3·3 years of Encke's Comet up to many thousands of years, while the inclinations of the orbits seem to be distributed entirely at random, so that roughly

as many comets travel retrograde as direct. Of some six hundred comets whose orbits are known with some certainty, about forty return regularly to the Sun, and these "short-period" comets have been studied extensively. They are usually faint objects, with orbits only slightly inclined to the ecliptic, and only one of them (Halley's Comet) travels in a retrograde orbit. These short-period comets, which are of particular interest to the meteor observer, are generally short-lived, and are seen at only a few returns. There are three notable exceptions: Halley's Comet, which has been recorded at each return (at roughly 75-year intervals) since 240 B.C.; Encke's Comet, seen every three years from 1819 to 1941; and Comet Pons-Winnecke, which has been seen at most returns since 1819.

A few cases are known in which several comets travel in orbits which, if not identical, are at least very similar. The most famous case of this is the group of four comets 1843 I, 1880 I, 1882 II, 1887 I. All four comets travelled in orbits of great size and high eccentricity, having small perihelion distance, similar inclinations and direction of the major-axis in space, and in all four cases the motion was retrograde. The phenomenon of comet-groups is exactly analogous to the behaviour of meteor-streams in travelling in an orbit similar to that of a comet.

The perihelion distance (q) of these bodies is of great importance, for it is clear that only those comets with a value of q small enough to bring them within visible distance of the earth will actually be seen by us, and there may be thousands of comets travelling in much larger orbits, and remaining invisible until some disturbance alters the orbit. It is even more important for a meteor-stream to have a suitable value of q, for the shower will only be visible in this case when the earth actually intersects the orbit of the stream, and this can only happen in the neighbourhood of the nodes. At the ascending node, the stream of particles is passing north across the plane of the earth's orbit, and the apparent radiant of the resulting meteors will lie in the southern hemisphere. In the same way, the intersection of the earth's orbit with that of the stream at a descending node, will give a radiant in the northern hemisphere of the sky.

Cometary orbits also show a wide range of eccentricities. It was formerly regarded as one of the characteristics of comets that their orbits were always very eccentric, but this point of view can no longer be held, since comets with quite small eccentricities are now known. Thus Comet Schwassmann-Wachmann (1) 1925 II has $e = 0{\cdot}135$ while Comet Oterma 1943 a has $e = 0{\cdot}144$, and the orbits, with their small inclinations, are quite indistinguishable from those of a minor planet. The only point of difference between the two classes of object is their appearance, for whereas a minor planet presents an almost stellar appearance in the telescope, a comet always appears as a hazy patch of light. This *coma* sometimes has a stellar nucleus, and as the comet approaches the Sun, it may develop a tail. In the case of a bright comet, a great deal of structure may be visible in the coma and in the tail, but many comets do not develop in this way.

§2. STRUCTURE

The head of a comet—the nucleus and its surrounding coma—obeys the law of gravitation, travelling in an orbit about the Sun; but the tail does not, for when a tail is formed, it always points away from the Sun. It is generally agreed that the tail consists of gases and minute particles, expelled from the head, and thereafter repelled from the Sun by radiation pressure. Although this

radiation pressure may well be one of the principal factors in the development of the tail, it is doubtful whether it is adequate to explain the shapes of the envelopes composing the coma, or the velocities of individual portions of the tail itself. The changes in brightness of a comet also cannot be explained in any simple way, for no one law has been found which can be used to express the visual magnitude of a comet in terms of its distance from the earth and from the Sun. Each individual comet has to be examined and an empirical law deduced to express the way in which its light varies (Vsessviatsky 1933, 1935, 1937). Sudden and unexplained outbursts of light are also noticed. Comet 1925 II, for instance, has a habit of producing outbursts of this kind, and in January 1946 it increased its brightness from the 17th to the 9th magnitude—a 1500-fold increase of light—in the space of a few days. Whatever may be the forces which are responsible for the creation and disruption of comets' tails, it is clear that the tail cannot be the origin of meteor-streams, for the tail is always directed away from the orbit, and the resulting momentum of the particles is quite different from that of the comet.

Both the head and the tail of a comet are completely transparent, and even small stars shine through them without loss of light. The comet of 1882, and Halley's Comet in 1910, passed between the Earth and the Sun, and nothing whatever was seen of them as they crossed the solar disc. The nucleus, however, must possess mass, since it obeys the law of gravitation, and the size of this nucleus may vary over wide limits. There seems little doubt that the nucleus of Comet Pons-Winnecke was of the order of 1 km. in diameter, while values up to 12 000 km. have been given for other comets. Actual measurement of the nucleus is a matter of some difficulty, and various assumptions are made in such work. There is every reason to believe that the nucleus consists of a cluster of meteoric bodies of various sizes, and the transparency and small mass support this view. The spectrum of sunlight is given by the head as it approaches the Sun (even the Fraunhofer lines are seen) and this suggests reflection from solid bodies, while the band and line spectra which are observed show the presence of just those elements which we know to be present in meteorites which have fallen upon the earth.

The actual mass of the nucleus must be very small. Lexell's Comet of 1770 passed very close to the satellites of Jupiter without disturbing them, and the same comet came within $2\frac{1}{2}$ million km. of the earth without causing a change of any kind (as, for example, a change in the length of the year). This, and similar facts, enable us to assess an upper limit to the mass of a comet, of the order of one millionth of the earth's mass, and this would suggest a mean density for the nucleus of the order 6×10^{-9}. A recent determination of the mass of Halley's Comet by Vorontsov-Velyaminov (1946) is based on measurements of the brightness of the Fraunhofer spectrum of the nucleus, and on deductions from the physical theory of spectra. He finds a mass of 3×10^{19} gm. (which is of the same order as previous estimates), but finds the nucleus to have a diameter of about 30 km., and hence to consist of a dense cluster of meteoric blocks, whose distances apart are comparable with their dimensions. Whether the nucleus consists of large blocks of this kind, or of smaller particles widely separated (Dubiago 1942) there is no doubt that the forces that maintain stability in the

nucleus are very remarkable, and gravitational attraction alone does not seem capable of accounting for this.

§3. DISRUPTION

In the case of a close approach of a comet to a major planet or to the Sun, it is to be expected that the ordinary Keplerian forces would be sufficient to disintegrate such a cluster of particles, or at least to elongate it in the direction of motion. Such an event was certainly witnessed in the case of the great comet of 1882, when the nucleus, after a particularly close approach to the Sun, became elongated into five separate nuclei "like a string of pearls". Yet even this enormous disturbance did not actually disintegrate the comet—it was still a comet, although with a multiple nucleus. Other cases are known in which a comet has passed through violent stresses of this kind without any apparent change. On the other hand, the actual disruption of a comet has been witnessed on several occasions, yet there has never been any clear understanding as to the cause of this. The classic case is that of Biela's Comet. In 1826, when its orbit was computed for the first time, it was found to have made two previous appearances. It was seen again in 1832, and in 1846, but on this last occasion, it divided into two parts which gradually separated. The twin comets appeared again in 1852, but have not been seen since. In 1872, however, as the Earth crossed the track of the lost comet, there was a fine shower of meteors from Andromeda—the Bielids or Andromedids. It is this event that has given rise to the idea that a meteor shower is caused by the disruption of a comet ; but the Bielids had given good displays of meteors on at least seven occasions before 1872 (as far back as 1741), and the same is true of the other major showers. The Leonids have been seen in November since 902 A.D., and there are records of the Perseids in August for over 1200 years, and of the Lyrids in April for 2500 years. Yet the comets which are associated with these showers were not discovered until the nineteenth century.

The disruption of Biela's Comet arose from unknown causes, for there were no major planets near enough to cause any disturbance. The shower of 1872 could not have arisen from the comet itself, the computed position of which was far from the earth at that time. Thus although a definite connection exists between comets and meteors, there is no direct evidence that meteors are formed by the disruption of a comet—indeed, all the evidence suggests that the disruption of a comet gives two or more comets, and not meteors. Disruptions of this kind, from whatever cause, may be responsible for the formation of the short-period comets from the thousands of comets that must lie beyond the reach of our instruments ; but meteors can only be formed from a comet when that body has lost all its gases, and the forces of cohesion no longer exist.

§4. PERTURBATIONS

It cannot be too strongly emphasized that it is Jupiter who is responsible for most of the perturbations of comets and meteor-streams. The disturbances caused by other planets are quite small in comparison with those set up by Jupiter, and theories of the capture of comets and meteor-streams by other planets must be discredited. The so-called "families of comets", consisting of comets whose aphelia all lie close to the orbit of one of the outer planets, do not, in fact, exist. It was shown by Russell (1920) that Jupiter alone could cause the formatiou

of such orbits, and examination of these "families" shows that the aphelia often do not lie near the planet which is supposed to be responsible. Halley's Comet, for instance, which is supposed to belong to Neptune's "family", does not approach that planet within 8 units.* It can, however, come within 1 unit of Jupiter. In order to force a comet into a new orbit in this way, an extremely close approach to Jupiter is necessary, and the perturbations caused at more normal distances, or by a smaller planet, will merely modify the orbit to a small extent.

These minor perturbations are of importance in the theory of meteor-streams. The effect of the attraction of one of the planets on a comet or other body may be resolved in three directions at right angles. Of these, the tangential component always causes a change in the size of the orbit, since in the velocity formula $V^2 = 2/r - 1/a$ (where V is the velocity at distance r from a central body in an ellipse of semi-major axis a) a force which tends to diminish V will cause a to decrease, and the body will travel in a smaller orbit. The normal component, acting in the plane of the orbit at right angles to the tangential component, causes a change in the eccentricity of the orbit and the direction of its major axis. The orthogonal component, perpendicular to the orbit-plane, is always present in any inclined orbit, and always has the same effect—it causes a regression of the line of nodes if the body is moving direct, and an advance of the line of nodes in a retrograde orbit.

This last point is of particular importance in identifying a meteor-shower with a comet. It is not merely sufficient to show that the shower travels in the same orbit, and that it recurs in the same period as that of the revolution of the comet about the Sun; it is also necessary to show that the date on which the shower occurs, which gives the position of the node, advances or retrogrades by the required amount. It was this fact that enabled J. C. Adams in 1867 to decide between the five possible orbits which H. A. Newton had postulated for the Leonids. Adams was able to show that only the 33-year period accounted for the advance of the node by one day in 70 years. Similarly the Andromedids were identified with Biela's Comet because they showed the same rapid regression of the node—about one day every 7 years.

In order to illustrate the effect of these perturbations on a comet, the elements of Comet Pons-Winnecke are given here for a number of its returns:

Date	Ω	i	e	q	Distance C – E	Computed radiant
1858	113·5°	10·8°	0·755	0·769	−0·231	
1886	104·1	14·5	0·726	0·885	−0·128	204° +45°
1915	99·4	18·3	0·701	0·972	−0·041	208 +54
1921	98·1	18·9	0·678	1·041	+0·030	205 +57
1939	96·8	20·1	0·670	1·102	+0·092	205 +59
1945	94·4	21·7	0·654	1·160	+0·150	207 +60

This is, perhaps, an extreme case, for this comet has an orbit whose aphelion lies near to the orbit of Jupiter, and since the period of the comet is almost exactly half that of the planet, there are considerable perturbations at each alternate revolution. The ascending node of the comet's orbit lies close to Jupiter's orbit at aphelion, and as a result of this, the tangential and orthogonal components reach considerable proportions over a period of several months. The

* The unit is the distance of the Earth from the Sun.

regression of the node (Ω), and the change in inclination (i) are caused by the orthogonal component, while the tangential component is responsible for increasing the size of the orbit. The resulting change in the distance of the comet's orbit from the earth is given in the table as the difference between the radii vectores at the node, in the sense Comet *minus* Earth; and the computed radiants of the associated meteor-shower are also given. There was a good shower in 1916, but the distance has now increased so rapidly that it seems unlikely that it will ever be seen again.

It will be clear that the extent of such perturbations will depend on the time during which the attractive forces are acting, and this will depend in its turn on the size of the orbit and its inclination. The time during which a body remains under the influence of a planet will be much shorter if the body travels in an orbit greatly inclined to the orbit of the planet, and, in the limit, the least effect will be observed if the body is travelling in the reverse direction. Thus comets such as Halley's, which have a retrograde motion, are much less disturbed than those which travel direct. It is this fact which accounts for the great age of the major showers, for the Leonids ($i=163°$) and Perseids ($i=115°$) travel in retrograde orbits, while the Lyrids ($i=80°$) are highly inclined to the ecliptic.

Apart from the effect of perturbations on a body, the combined effect of the Sun's gravitational attraction and radiation pressure would cause it to move in a variable ellipse (Fessenkoff 1946). A more complete investigation, including the effects of relativity on the moving body (Robertson 1937), suggest that the combined effect of these forces is to cause a drag on the body, which therefore describes a slow spiral into the Sun. As a result of this, the neighbourhood of the Sun would be cleared of all small particles in the course of astronomical time.

The effect of the earth on a meteor-stream is necessarily small. Those meteors which come within the influence of the Earth's attraction are, however, forced to follow a hyperbolic orbit about the Earth, so that any of these meteors which are observed to fall through the atmosphere will have radiants which are displaced from the real position. The effect is always such as to raise the observed radiant towards the zenith, and a correction for this "zenith attraction" must always be considered. The maximum value of this correction is reached when the meteor is travelling in the same direction as the earth, from a radiant near the horizon, and it may then amount to 17°. A second smaller correction is also applied to allow for diurnal aberration, especially for radiants near the pole.

§5. METEOR-STREAMS

The effect of perturbations on a swarm of particles, as distinct from a single body, will be to spread the swarm out into a flat ring round the Sun, and subsequent perturbations will cause the particles to become unevenly distributed round the orbit. The forces which bind together the head of a comet are not present here, and each individual particle behaves independently. Nothing could be farther from the truth than the popular idea of a "bicycle-tube" structure for such a swarm, for a uniform distribution of the dust around the orbit is impossible. When there is some form of distribution right round the orbit, there will be a shower every year, at the place and time at which the earth crosses the orbit. Such a distribution will presumably occur only in the case of a swarm of great age, and in the case of streams of more recent formation,

there will be showers only at certain places in the orbit. The dust will not, however, follow behind the comet, as is popularly supposed. This may be seen by reference to any of the major showers, but is perhaps more convincing if tabulated for a recent shower, that associated with Comet Giacobini-Zinner :

Date	e	q	Distance C – E	Earth reaches node :
1900	0·733	0·9319	−0·0619	
1913	0·720	0·9759	−0·0181	
1926	0·717	0·9947	+0·0005	70 days before comet : shower
1933	0·716	0·9994	+0·0054	80 days after comet : shower
1939	0·717	0·9955	+0·00125	136 days before comet : no shower
1946	0·717	0·9957	+0·00142	15 days after comet : shower

The prediction of a shower of meteors from any comet is clearly a difficult matter, since the meteoric dust lies in patches around the orbit, or to be more precise, in a series of neighbouring orbits, which bear some relation to that of the comet.

We can estimate the extent of the distribution of this dust in a simple manner. Taking the earth as a sphere of radius 6400 km. travelling at 30 km/sec., and accepting as a generous estimate that 10^8 meteors fall on the Earth every 24 hours, we find there will be one particle in every $3{\cdot}3 \times 10^6$ km^3 of space, corresponding to one particle every 130 km. In a shower such as the Perseids, the packing will be closer than this, but even in an unusual shower, like that of the Leonids in 1833, when the hourly rate rose to 35 000, the distance between the particles must have been of the order of 20 km. Thus there is a very great difference between the structure of a meteor-shower and that of the nucleus of a comet.

In the case of the Perseids, we can estimate the width of the stream from the fact that the shower lasts about 21 days, during which time the Earth will have travelled 5×10^4 km. The stream crosses the Earth's orbit at an angle of about 14° so that its width must be about $1{\cdot}3 \times 10^7$ km. As it is unlikely that the Earth makes a central passage through the stream, this figure is a lower limit. The amount of matter removed from this stream at each passage of the Earth is negligible in comparison with the whole stream, and the failure of a shower to appear as expected, is not due to its losses from capture by the Earth, but to perturbations or to the non-uniform distribution of the dust in the orbit. Another factor which has only recently become important is that the shower may approach the Earth from the sunlit side, so that the radiant is only present in daylight. It is this type of shower which is presumably responsible for the radio-echoes detected in the spring of 1947 (Lovell).

The width of the Perseid stream ensures that the radiant shall be active for several nights in succession, but in the course of 24 hours, the Earth has traversed about one degree of longitude, and hence the angle at which the meteors enter the atmosphere has changed by about the same amount. The movement of a radiant by about one degree in longitude on successive nights is thus a further proof of the presence of a meteor stream, but this change may be seriously modified in the case of a radiant near the ecliptic.

There are thus four criteria by which we may judge the association of a meteor-shower with a comet, namely (i) similarity of orbit, with close approach to the Earth ; (ii) periodic return of the shower ; (iii) advance or regression of the node ; (iv) daily movement of the radiant. Complete agreement on all four

counts is rare, and the known cases in which agreement may be said to be highly probable are comprised in the following list :

Shower	Comet	Computed date	Computed radiant	Distance C – E
Lyrids	1861 I	Apr. 21	271° +34°	−0·002
June Draconids	Pons-Winnecke	June 30	208 +54	−0·042
Perseids	1862 III	Aug. 11	45 +58	+0·010
October Draconids	Giacobini-Zinner	Oct. 10	262 +54	+0·004
Leonids	1866 I	Nov. 15	151 +23	−0·052
Andromedids	Biela	Nov. 30	23 +44	−0·018

The dates given in this table have been computed from the original cometary orbits, and it is interesting to note that the change in the nodes has, in the course of time, inverted the order of the last two showers. The Andromedids now occur on November 14, and the Leonids on November 16.

In addition to the showers given in the table above, the Taurids are shown by Whipple (1940) to have a common origin with Encke's Comet. Other writers (Olivier 1925, 1930 ; Hoffmeister 1937) definitely assign the May Aquarids to Halley's Comet. The idea of a common origin for a comet and its associated meteor stream is attractive, for it cannot be said that in these two cases the present orbit of the comet could give rise to these showers. Both comets, having small inclinations, approach the Earth's orbit more closely at some distance from the node than at the node itself, but even so the distances concerned are extraordinarily large—Halley 0·15, Encke 0·19 unit. Some limit must be placed on any definition of " a close approach ", and the distances given in the table above would suggest that this limit is quite small. There are 35 comets which can come within 0·05 unit of the Earth's orbit, while 60 come within 0·1 unit. Among these there are no obvious accordances with meteor showers apart from those tabulated above.

There remain a number of showers each of which appears regularly, but for which the parent comet remains unknown, if, indeed, there ever was a parent comet. For it must be admitted that there is no *direct* evidence of the disruption of comets to form meteor-streams, and the number of accordances between showers and comets is very small. The most that can be claimed is that there are a few cases of accordance, and it seems plausible to suggest that comets and meteors have a common origin. It is to be hoped that the new radio-echo methods will bring us much new information about meteors, especially those that fall on the sunlit side of the Earth.

REFERENCES

DUBIAGO, A. D., 1942, *Russian Astr. J.*, **19**, 14.

FESSENKOFF, B., 1946, *Russian Astr. J.*, **23**, 353.

HOFFMEISTER, C., 1937, *Die Meteore* (Leipzig).

LOVELL, A. C. B., 1947, *Brit. Astr. Assocn. Circular* 282.

MERTON, G., 1947, *Observatory*, **67**, 117.

OLIVIER, C. P., 1925, *Meteors* (Baltimore : Williams and Wilkins) ; 1930, *Comets* (London : Baillière, Tindall and Cox.)

ROBERTSON, H. P., 1937, *Mon. Not. R. Astr. Soc.*, **97**, 423.

RUSSELL, H. N., 1920, *Astr. J.*, **33**, 49.

STRÖMGREN, E., 1914, *Pub. Copenhagen Obs.*, No. 19.

VORONTSOV-VELYAMINOV, B., 1946, *Astrophys. J.*, **104**, 226.

VSESSVIATSKY, S., 1933, *Russian Astr. J.*, **10**, 327 ; 1935, *Ibid.*, **12**, 243 ; 1937, *Ibid.*, **14**, 480.

WHIPPLE, F. L., 1940, *Proc. Amer. Phil. Soc.*, **83**, 711 ; 1947, *Ibid.*, **91**, 189.

SPECTROSCOPIC OBSERVATIONS ON COMETS, METEORS AND METEORITES

By R. W. B. PEARSE,
Imperial College, London

§1. INTRODUCTION

THOUGH the origin of their constituent material and the scene of its aggregation are both obscure there appears to be no doubt that comets, meteors and meteorites are intimately related. A comet is regarded as a diffuse collection of solid bodies, mostly very small in size, describing a closed orbit about the Sun. In the course of time these bodies become scattered along their orbit, and, if this intersects the orbit of the Earth, they may periodically encounter the earth, giving rise to the phenomenon of meteor showers. The smaller particles are entirely consumed in the Earth's atmosphere, but larger bodies, which may reach several tons in weight, reach the Earth's surface and are to be identified with meteorites.

Three opportunities thus arise for spectroscopic observations on this interplanetary material. First the light received from the comet as it passes near the Sun may be analysed, then that emitted by the particles as they become incandescent in the Earth's atmosphere and finally the material of the meteorites may be subjected to laboratory investigation. The sections which follow give a brief survey of results which have been obtained from these three fields of research.

§2. THE SPECTRA OF COMETS

At the greatest distances from the Sun at which they can be observed comets can scarcely be distinguished from the neighbouring background stars, except by their motion relative to these stars. But as they come nearer the Sun they are seen to become diffuse. In the brightest comets three regions are usually distinguished : The *nucleus*, a small bright region, the *coma*, a diffuse luminous region enveloping the nucleus, and the *tail*, a long train of weaker luminosity stretching for millions of miles in a direction away from the Sun. These three regions call for separate spectrographic study.

Two methods of observing the spectrum are available, each of which has its special advantages. In the *objective prism method* a prism is placed before the lens which is used to photograph the comet. This disperses the light so that a separate complete image of the comet is formed for each wavelength emitted. This method has the advantages of showing the distribution of light of each wavelength throughout the comet and of giving a result for the fainter regions of the comet.

In the *slit spectrograph method* an image of the comet produced by a telescope is thrown on the slit of a complete spectrograph. This method gives better definition of the spectral lines, but the result refers only to a narrow strip of the comet covered by the slit.

The early spectrographic studies of comets soon established that the spectrum

of the head consisted partly of continuum and partly of discrete bands. The continuum is certainly in the main due to scattered sunlight and shows Fraunhofer lines, but part has been claimed to have another origin and to show a different variation of intensity with heliocentric distance (Bobrovnikoff 1942). The strongest bands were identified as belonging to the spectra of the diatomic molecules C_2, CH and CN. Objective prism spectra showed that the faint light of the tail also consisted of bands. These were found to be mainly due to CO^+ though some bands of N_2^+ were also found.

In these investigations the resolving power was not great. The recent employment of quartz spectrographs of better resolving power has both extended the range of observation and given improved detail in the spectra. This again has led to renewed interest in the subject of cometary spectra. Specially notable is the series of spectrograms of Comet Cunningham, 1940 c, obtained at the McDonald Observatory (Swings, Elvey and Babcock 1941). These revealed clearly for the first time the presence of bands due to OH and NH. The long unidentified group of bands in the region of λ4050 received special study and were subsequently duplicated in the laboratory by Herzberg (1942) in the spectrum of an interrupted discharge through rapidly streaming CH_4. The bands most probably originate in the molecule CH_2. Another triatomic molecule, NH_2, has been considered by Swings, McKellar and Minkowski (1943) to account for a number of emission features in the visual region. Swings (1943) also considers it probable that OH^+ and CH^+ are present, the latter extending into the tail. Atomic lines do not as a rule appear in the spectrum, but occasionally, on very close approach to the Sun, such lines as the D lines of Na have been observed in emission. Details of the bands identified are given in table 1.

Table 1. Bands in the spectra of comets

Molecule	System	Occurrence	Extent
C_2	Swan, $B\,^3\Pi \rightarrow A\,^3\Pi$	Head	Five sequences : (0,2) λ 6191 . . . (0,1) λ 5635 . . . (0,0) λ 5165 . . . (1,0) λ 4737 . . . (2,0) λ 4382 . . .
CH	$A\,^2\Delta \rightarrow X\,^2\Pi$	Head near nucleus	(0,0) λ 4300
	$B\,^2\Sigma \rightarrow X\,^2\Pi$	Head near nucleus	(0,0) λ 3900
	$C\,^2\Sigma^+ \rightarrow X\,^2\Pi$?	Head near nucleus	(0,0) λ 3145 falls on (1,1) band of OH
CH^+	$^1\Pi \rightarrow {}^1\Sigma$	Tail	Lines near λ 4240 and λ 3960
CH_2	λ 4050 group	Near nucleus	Lines between λ 3980–λ 4080
CN	Violet, $^2\Sigma^+ \rightarrow {}^2\Sigma^+$	Head	Four sequences : (0.2) λ 4604 (0,1) λ 4216 (0,0) λ 3883 (1,0) λ 3590
	Red $^2\Pi \rightarrow {}^2\Sigma^+$	Head?	Identification uncertain
CO^+	Comet-tail $^2\Pi \rightarrow {}^2\Sigma$	Tail	Double-headed bands : λλ 3269, 3436, 3586, 3782, 4003, 4257, 4549
N_2^+	$^2\Sigma \rightarrow {}^2\Sigma$	Tail	(0,0) λ 3914
NH	$^3\Pi \rightarrow {}^3\Sigma$	Head	(0,0) strong lines near λ 3360
NH_2	α-bands	Head near nucleus	Many coincidences between λ 5100–λ 6700
OH	$^2\Sigma^+ \rightarrow {}^2\Pi$	Head	(0,0) λ 3078–λ 3100 (1,1) weak lines λ 3134–λ 3152 (2,2) two weak lines?
OH^+	$^3\Pi \rightarrow {}^3\Sigma$	Head	Possible contribution from (0,0) band at λ3560–λ3600

It is notable that all the band systems observed involve the ground state of the molecule. This is consistent with the view that the emission is due to fluorescence excited by the Sun. Confirmation has been obtained in a very interesting way by McKellar (1942) through quantitative spectro-photometric work on the rotational structure of the bands, following a suggestion due to Swings. Briefly, if the upper rotational levels are populated by absorption of sunlight, then the population will depend on the contour of the solar spectrum in the absorption region. The distribution of intensity in the rotational structure in the succeeding emission will therefore in some measure reflect the solar contour. The effect may be very marked if the region contains prominent Fraunhofer lines, moreover, through Doppler displacement, the relative velocity of the comet with respect to the Sun may also influence the absorption. McKellar has shown that this mechanism accounts for the structural features of the CN bands in considerable detail.

§3. SPECTRA OF METEORS

Meteors provide a difficult source of light for spectroscopic investigation. Their appearance is sudden, of brief duration and very random in time and direction. The systematic study of their spectra was begun by Millman (1932, 1933) who, collecting the recorded data then available, found details of only nine observations. These had all been obtained by chance with objective prism apparatus in use for other purposes. However, even so, it was possible to draw some general conclusions regarding the emission from these bodies.

The dispersion was of course necessarily small, but nevertheless it could be determined that in each case the spectrum consisted of bright atomic lines and that there was marked similarity between the spectra of the various meteors. Lines of a number of elements could be identified. Thus Millman found lines of Fe to be present in all nine of the spectra and lines of Ca^+ in six of them. Other lines were definitely attributed to Mg, Mn, Cr and Al. The evidence for Si was inconclusive. Beyond the photographic range visual observations showed the yellow lines of Na. All of the lines were found to be low excitation lines of the atoms concerned. In a majority of the spectra the most intense lines were the H and K lines of the ion Ca^+.

Interest was stimulated by these results, and in a later paper Millman (1935) was able to consider a greater number of spectra obtained during the intervening period as the result of more organized observation. The element Ni was added to the list of those identified and the spectra were divided into two groups according to the degree of prominence of the Ca^+ lines. In type Y the H and K lines are the most prominent feature, in type Z they are markedly absent.

The spectrum, consisting as it does of lines, must originate in material in the state of vapour. Meteors are therefore presumably surrounded by layers of hot gas generated by frictional heating as they plough their way through the atmosphere. Possibly combustion also contributes something to the heat generated, but so far no radiation from molecules has been reported. As described in the following section study of meteorites shows two contrasting types, the iron meteorites and the stone meteorites. The former contain very little Ca while

the latter contain more than one per cent. It is therefore probable that type Y spectra are to be attributed to stones. Meteors recognized as belonging to showers have given, so far, spectra of type Y, and so are considered to be mainly stone. Sporadic meteors are more evenly divided. Measurements of the relative intensities of the Fe lines have been used to make estimates of the temperature existing in the vapour. They give results comparable with those obtained with furnace excitation in the temperature range 1700°–3300° A.

The brightest meteors leave trains which remain luminous for a considerable period, sometimes extending over more than an hour. This luminosity appears to occur at heights of the order of eighty kilometres. So far no one has been able to obtain a spectrum of this residual glow.

Since the chance of securing a spectrum is much increased at the times of meteor showers, spectroscopic studies may be usefully added to the programme of shower observations. Millman (1940 a, b) has given accounts of such studies for the Perseid meteor showers of 1939 and1940.

§4. SPECTROSCOPIC STUDIES OF METEORITES

While bearing in mind the possibility of some differences in composition, particularly of the outer layers, between a meteoric body fresh from its travel in the more distant parts of its orbit and one which has passed through the Earth's atmosphere and perhaps reposed for a considerable period on the Earth's surface, one may nevertheless expect important results bearing on the above conclusions regarding comets and meteors to be obtained from laboratory studies of meteorites.

Laboratory spectroscopic investigations of meteorites lead to the detection of a far greater number of elements than appear in the spectra of comets and meteors. Considerations of abundance and ease of volatilization, however, lead to the conclusion that these spectra are not inconsistent with the view that comets and meteors are composed of the same material as meteorites. As examples of spectroscopic analysis of meteorites, recent work by King (1936 a, b) and by Johnson and Norman (1943) may be cited.

King's results were obtained from four iron meteorites, eight stone meteorites and one graphite nodule. Altogether, apart from carbon, thirty-one elements were observed. The average compositions found for iron and stone meteorites were given separately. Johnson and Norman studied eighteen meteorites of which details are given in their paper. The spectroscopic results for each meteorite are given separately and show considerable variation in the percentage composition in passing from sample to sample. A search was made for each of sixty-nine elements and of these forty were detected. Twelve of these had not been reported by King, but on the other hand, King reported three not found by Johnson and Norman. These spectroscopic results can be supplemented by chemical analyses such as those of I. and W. Noddack (1930). These provide data for the non-metallic elements not observed in the spectroscopic studies. In all, these authors give twenty-six elements as present in measurable amount and a further thirty-seven as present in traces. A list of the more abundant elements is shown in table 2.

Table 2. Percentage of more abundant elements in meteorites *

Iron				Stone							
Fe	90·8	S	0·04	O	36·3	Ca	1·7	Cr	0·32	Ti	0·11
Ni	8·5	C	0·03	Fe	23·3	Al	1·5	Mn	0·24	P	0·11
Co	0·59	Cu	0·02	Si	18·1	Ni	1·5	Co	0·17	Cl	0·08
P	0·17	Cr	0·01	Mg	13·7	Na	0·65	K	0·17	Cu	0·01
				S	1·8			C	0·15		

* The figures represent averages. Individual meteorites show considerable variations.

For comparison with the behaviour of comets it is a knowledge of the gases which may be liberated from meteorites which is the main interest. By pulverization and heating *in vacuo* H_2, CO, CO_2, CH_4, N_2, H_2S and SO_2 have been obtained in considerable quantity according to Merrill (1926) and H. von Klüber (1931). The proportion of the different constituents of the gas for iron meteorites is different from that for stone meteorites.

Attempts have been made to determine the ages of meteorites, that is, the time which has elapsed since their solidification, by the method of radioactive decay which has been used to determine the ages of rocks. The amounts of He, U and Th contained in the meteorite may be determined and the time necessary for the liberation of the He from the U and Th may be calculated. Paneth (1939) at first neglecting the Th content, found for twenty-nine meteorites ages ranging from 100 to 2800 million years. For three meteorites where the Th was determined it was found to be important. The ages obtained were 30, 500 and 1000 million years

Later Arrol, Jacobi and Paneth (1942) extended the observations to a great many more meteorites. From these results Bauer (1947) draws attention to a remarkable variation of the concentration of He with the mass of the original meteorite, of such kind that large He concentrations appear only in relatively small masses and large masses all have relatively small He content. Bauer suggests that cosmic radiation has produced extra helium in the outer layers and makes an estimate that cosmic ray disruption alone would produce the maximum He content in less time than the present assigned age.

REFERENCES

Arrol, W. J., Jacobi, R. B., and Paneth, F. A., 1942, *Nature, Lond.*, **149**, 235.

Bauer, C. A., 1947, *Phys. Rev.*, **72**, 354.

Bobrovnikoff, N. T., 1942, *Rev. Mod. Phys.*, **14**, 164.

Herzberg, G., 1942, *Astrophys. J.*, **96**, 314.

Johnson, W. W. A., and Norman, D. F., 1943, *Astrophys. J.*, **97**, 46.

King, A. S., 1936 a, *Astrophys. J.*, **84**, 507 ; 1936 b, *Pop. Astron.*, **44**, 507.

von Klüber, H., 1931, *Das Vorkommen der chemischen Elemente im Kosmos* (Leipzig).

McKellar, A., 1942, *Rev. Mod. Phys.*, **14**, 179.

Merrill, G. P., 1926, *Proc. Amer. Phil. Soc.*, **65**, 119.

Millman, P. M., 1932, *Harvard Annals*, **82**, No. 6 ; 1933, *Harvard Bull.*, 891.; 1935, *Harvard Annals*, **82**, No. 7 ; 1940 a, *Communication David Dunlap Obs.*, No. 5 ; 1940 b, *Ibid.*, No. 7.

Noddack, W., and Noddack, I., 1930, *Naturwissenschaften*, **18**, 757.

Paneth, F. A., 1939, *Occ. Notes R. Astr. Soc.*, No. **5**, 57.

Swings, P., 1943, *Mon. Not. R. Astr. Soc.*, **103**, 86.

Swings, P., Elvey, C. T., and Babcock, H. W., 1941, *Astrophys. J.*, **94**, 320.

Swings, P., McKellar, A., and Minkowski, R., 1943, *Astrophys. J.*, **98**, 142.

METEORIC IONIZATION AND IONOSPHERIC ABNORMALITIES

By A. C. B. LOVELL,
Physical Laboratories, University of Manchester

INTRODUCTION

WHEN a meteor enters the Earth's atmosphere it creates a trail of ionization at a height of about 100 km. above the surface of the Earth. Under suitable conditions this ionization scatters back sufficient energy from a radio wave to give a transient response on a sensitive receiver. The recent developments of such radio-wave technique have provided an instrument of greater sensitivity than the human eye for the investigation of meteors by day or night which is beginning to yield new data about the physics of the atmosphere at these altitudes. There is also reliable experimental evidence that the general influence of meteoric ionization is one of the primary causes of the longer period abnormalities in the E region of the ionosphere. This Report is divided into two main sections. Part I (§§ 1–6) summarizes the historical development of the work on the abnormalities in the E-region ionization and discusses the present day views on the origin of sporadic-E ionization. Part II (§§ 7–8) deals with the contemporary work on the transient responses due to meteoric ionization and the application of this work to the astronomy of meteors and the physics of the high atmosphere.

PART I. ABNORMAL EFFECTS IN THE E REGION AND THEIR RELATION TO METEORS

§ 1. THE GENERAL FEATURES OF THE STRUCTURE OF THE IONOSPHERE

The main features of the structure of the ionosphere were elucidated by studying the reflection of radio waves from the upper atmosphere, using either the frequency-modulation method of Appleton and Barnett (1925 a, b) or the pulse method of Breit and Tuve (1925 a, b, 1926). The existence of two principal ionized regions—the E region at an altitude of 120 km. and the F region at an altitude of some 250 km.—was demonstrated by Appleton (1927 a, 1930), Appleton and Ratcliffe (1930), and Appleton and Green (1930) in Great Britain, using the frequency-modulation method and by Schafer and Goodall (1931) in America, using the pulse technique. A great amount of detail has since been obtained about these ionospheric layers and comprehensive reviews of the subject have been given by Appleton (1935, 1939) and by Mimno (1937). For the purpose of this report it is sufficient to mention that under normal conditions the electronic density in the E region shows a marked solar control reaching a maximum at noon and a minimum (due to recombination and attachment) during the night. For example, in Great Britain at noon on a summer day the density in the E region is approximately $1{\cdot}1 \times 10^5$ electrons per cm^3, decreasing to about $7{\cdot}9 \times 10^3$ electrons per cm^3 by midnight. The equivalent critical frequencies for radio waves are 3 Mc/s. and 0·8 Mc/s. respectively. Waves

of higher frequency penetrate the E region and are not reflected. The mean electronic density in the F region is about 6×10^5 electrons per cm^3 corresponding to a critical frequency of 7·0 Mc/s.

§2. THE DISCOVERY OF ABNORMAL EFFECTS IN THE E REGION

Evidence that the ionization in the E region did not decrease in a regular manner during the night was obtained by Heising in 1928. Heising used the pulse technique on 2·7 and 5·2 Mc/s. to study the heights of the ionized layers, and after sunset often found sudden decreases in height of the E layer corresponding to sudden increases in the electron density in the layer. His results are illustrated in figure 1, which shows the normal tendency of the ionization to decrease in the E layer at night with the occasional sudden increases. Heising says: "It would appear as though great masses of electrons are tossed into the atmosphere rather quickly and that as a result the level drops with accompanying turbulence and variation in density near the lower edge". Heising concluded that the ionizing conditions after sunset were not constant, or varying in a regular manner, and considered that particles from the Sun might be the cause of the irregularities. Further evidence of abnormalities was given by Eckersley (1929) using a different technique. He studied the propagation of radio waves in the frequency range 6 to 20 Mc/s. Under conditions when the skip zone effect was present, due to penetration of E but reflection from F, Eckersley found that signals were frequently scattered into the skip zone. He attributed this effect to the occurrence of dense ionic clouds below 125 km. but did not discuss their origin. Abnormal nocturnal effects were also reported by Appleton (1930) and by Appleton and Naismith (1932) who concluded that "either the recombination of ions is prevented or there is some ionizing agent present which can influence the dark side of the earth". Schafer and Goodall (1931) also found rapid minute to minute variations during the night which they regarded as indicating the presence of some ionizing agency other than the Sun.

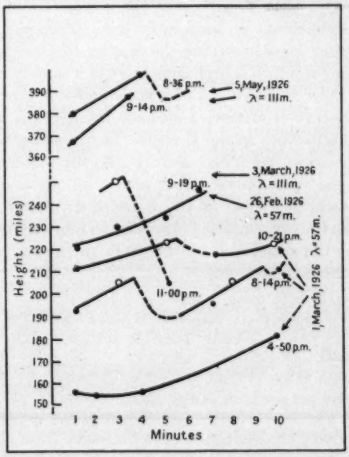


Figure 1. Heising's results on variation of height of the E region with time. Circles and dots are observations. Circles are observations in which multiple reflections occurred. (*From* R. A. Heising, *Proc. I.R.E.*, 1928, **16**, 75.)

§3. EARLY VIEWS ON METEORS AS THE CAUSE OF ABNORMAL EFFECTS IN THE E REGION

The suggestion that meteors might cause sufficient disturbance in the E region to affect the propagation of radio waves was first made by Nagaoka (1929, 1931). Nagaoka considered that a meteor would sweep away the electrons in its path, and although it would ionize the air itself the number of electrons produced would be small compared with the number present before the passage of the meteor.

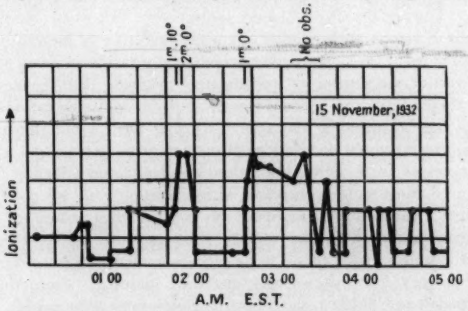


Figure 2. Transient increases in ionization of the E region associated with meteors passing overhead obtained by Skellett, Schafer and Goodall on 15 Nov. 1932.
(*From* Skellett, *Proc. I.R.E.*, 1935, **23**, 132.)

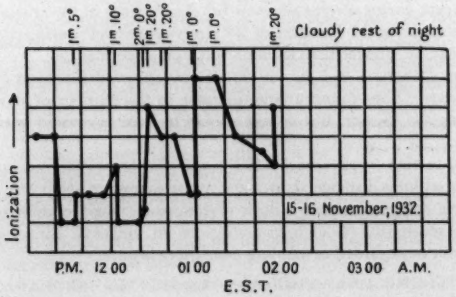


Figure 3. Transient increases in ionization of the E region associated with meteors passing overhead obtained by Skellett, Schafer and Goodall on 15–16 Nov. 1932.
(*From* Skellett, *Proc. I.R.E.*, 1935, **23**, 144.)

Thus, in the track of the meteor there would be fewer electrons than in the surrounding air and this would cause abrupt changes in the refractive index for an incident radio wave.

The first suggestion that the actual ionization due to meteors might affect the conditions in the E region was made by Skellett in 1931, who pointed out that Heising's results might be explained in this manner. Skellett's theoretical estimates of the amount of ionization produced by meteors, which he developed

further in 1932, were based on the theory of Maris (1929) and are now believed to be in error by an order of magnitude (Herlofson 1948). Nevertheless, the experimental work carried out by Schafer and Goodall (1932 a, b) during the Leonid shower of 1931, and by Skellett (1935) in conjunction with Schafer and Goodall, produced conclusive evidence that meteoric ionization was responsible for some of the night-time E-region abnormalities. These workers used the pulse method and observed on frequencies between 1·6 and 6·4 Mc/s. in rapid succession. During the 1931 Leonid shower very disturbed conditions were found in the E region with sudden abnormal increases of ionization, reaching a peak on the night of 16–17 November, which was known from visual observations to be the peak of the shower. Unfortunately magnetic disturbances were also present on these nights and hence it was not possible to decide definitely that the effects were due to meteoric ionization. During the Leonid shower of 1932, however, successful visual correlations of meteors passing overhead with sudden transient increases in E-region ionization were obtained (Skellett 1935). In figures 2 and 3 these sudden increases in ionization are shown correlated with the passage of meteors for the nights of 14–15 November and 15–16 November 1932. During nights when no meteor showers were active, these sudden increases in ionization were found to be rare—a similar plot given for the night of 19–20 February 1929, shows only one such increase. From this work of Schafer, Goodall and Skellett, the following conclusions were made : (*a*) night-time increases in E-region ionization were most marked during meteor showers, (*b*) for all major increases of ionization a meteor was observed to pass nearly overhead, (*c*) the intermittent reflections, lasting only for a period of seconds, were rare except during meteor showers, (*d*) observations of the critical frequency gave the ionization in the E region as 10^6 electrons per cm^3 during the maximum of the Leonids, i.e. greater than its noon value on a summer day.

In Japan, Minohara and Ito (1933) also investigated the effect of the 1932 Leonid shower. They found a large increase in the number of short duration echoes during the shower, but do not appear to have associated specific echoes and meteors.

Previous to this work Quäck (1931) and Pickard (1931) investigated the disturbances on long distance short-wave transmissions to see if any connection with meteoric showers existed. Both investigators found correlations, but in view of the sensitivity of such transmissions to magnetic disturbances their analysis cannot be regarded as yielding conclusive results.

§4. THE LONG DURATION ABNORMALITIES—SPORADIC E

The combined work of Skellett, Schafer and Goodall referred to above showed that the passage of a meteor in the Earth's atmosphere could give rise to a transient radio echo from the vicinity of the E region. Their results did not show, however, precisely what part the meteoric ionization played in the abnormal increases of long duration, and from their experimental data it does not seem possible to go beyond their general conclusion that " night-time increases in E-region ionization were most marked during meteor showers ".

In the years following the discovery of the long duration abnormalities there was a great deal of speculation as to the cause of the ionization increases, and it is convenient to consider the various suggestions under separate headings.

(i) *Thunderstorms and Barometric Effects*

Between May 1931 and June 1932 Ranzi (1932) made daily observations of the abnormal increases in E-region ionization and concluded that they were accompanied by particular isobaric situations, characterized by the presence of low pressure areas at the place of observation or to the north of it. C. T. R. Wilson had already pointed out some years previously (1925) that thunderstorms could cause ionization in the upper atmosphere and this, in conjunction with Ranzi's observations, led Appleton and Naismith (1933) to seek for a correlation between thunderstorms and the abnormalities found in their own observations. Between 1 January and 31 July 1932, the correlation coefficient between ionization density and a " thunderstorm index ", based on observations of atmospherics, was found to be 0·75. Ratcliffe and White (1933 a) obtained more evidence in support, stating " Our results entirely confirm his (Ranzi's) observations and seem to show an even closer correlation with thunderstorms and heavy rainfall ". Both Appleton and Naismith, and Ratcliffe and White, had shown that the agency responsible for the abnormal increases affected the E region only and not the ionization in the F region.

The thunderstorm hypothesis received further support from Appleton, Watson-Watt and Ratcliffe at a Royal Society Discussion in 1933. Ratcliffe and White (1934) found that the nocturnal increases occurred at a constant height of 105 km. and also that daytime sporadic echoes lasting for an hour or two occurred at the same height. The analysis of their data showed that 74% of the " thunderstorm days " showed nocturnal E-region abnormalities, whereas the effect only occurred on 46% of the days with no thunder.

Watson-Watt, Wilkins and Bowen (1937) also concluded that thunderstorms were responsible for maintaining the ionization at lower reflecting levels in the atmosphere, but the existence of these lower levels was not subsequently verified.

Bhar and Syam (1937) from measurements made in Calcutta, found a marked correlation of " abnormal E " with thunderstorms. In Australia, Martyn and Pulley (1936) found a correlation between the occurrence of nocturnal abnormal E and the barometric pressure observed on the ground the following morning. Kirby and Judson (1935 a, b) and Berkner and Wells (1937) found no correlation of abnormal E with thunderstorms, and finally Best, Farmer and Ratcliffe (1938) concluded that the evidence for the relation of abnormal E with thunderstorms was insufficient, and also that the barometric effect found by Martyn and Pulley did not exist in Great Britain.

After this date the idea that thunderstorms or barometric effects played a significant rôle in the abnormal increases of ionization in the E region was discarded. Because of uncertainty in assessment of a thunderstorm index and of the areas over which storms can be considered effective, such associations were difficult to interpret and the accumulation of statistics over a period of some years was necessary before the original ideas could be disproved.

(ii) *Magnetic Storms and Aurorae*

In 1929 Hafstad and Tuve made observations during a magnetic storm and found a very marked change in height of the E region and in the nature of the echo pattern. The possibility of a connection between the abnormal-E ionization

and magnetic activity was discussed by Appleton and Naismith in 1932. In 1933 these authors found a frequent but not invariable connection, but always an increase in E-region ionization during a big magnetic storm. During the Polar Year Expedition of 1932, this connection was investigated in detail. A preliminary account was given by Appleton, Naismith and Builder (1933), and the full account by Appleton, Naismith and Ingram (1937). Although the work showed a clear connection between magnetic activity and abnormal E ionization, the relation between the two was complicated. The general view was that the incidence of charged solar particles on the upper atmosphere would cause an increase in ionization and hence abnormal ionization would be expected to occur at the same time as magnetic storms and auroral activity. On occasions when the magnetic activity was particularly intense, however, the E echo often disappeared. To account for this fact it was suggested that when the solar particles were particularly numerous and penetrating, they would also ionize at a lower level in the atmosphere and thus form an absorbing fringe. Appleton, Naismith and Ingram include in their paper a graph plotted at a later date by Harang. This is reproduced in figure 4, and shows in a striking manner the

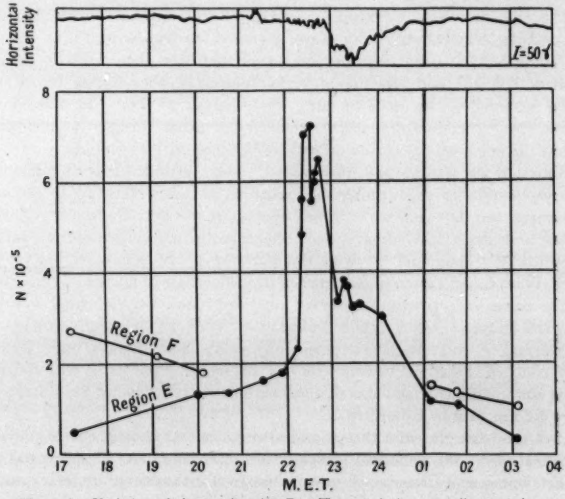


Figure 4. Variation of abnormal region E at Tromsö during a small magnetic disturbance, 22–23 March 1935.
(*From* Appleton, Naismith and Ingram, *Phil. Trans.*, 1937, **236**, 191.)

connection between the abnormal increase of ionization and a small magnetic disturbance.

The results of the Polar Year Expedition showed a definite connection between magnetic disturbances and abnormal-E ionization, but the complications of the phenomena were such that many subsequent workers disagreed. Thus Ratcliffe

and White (1934), Mimno and Wang (1933), Mimno (1937), and Kirby, Gilliland, Smith and Reymer (1936) obtained evidence in support of the Polar Year results, but Bhar and Syam (1937), Kirby and Judson (1935 a, b) and Maeda, Konomi and Isagawa (1935) found no correlation. Kirby, Berkner and Stuart (1934) found the abnormalities both on magnetically disturbed and on quiet days, and Berkner and Wells (1937) found that abnormal E was related to sharp magnetic discontinuities and auroral activity.

Radio " fade-outs " on short-wave circuits are now known to precede, but to be associated with such abnormalities. Dellinger drew attention to the sudden occurrence of this phenomenon in 1935, but the general association of disturbed propagation conditions with solar and magnetic disturbances had been discussed previously by Appleton (1927 b), Plendl (1932) and Mögel (1930, 1932). These effects appear to be associated with the " no echo " conditions found in the Polar Year work, referred to above, and are caused by particles from the Sun ionizing at a level below the E region, and hence forming an absorbing layer which prevents the normal reflection of radio waves at the E layer. More recent work by Appleton and Hey (1946) has demonstrated the connection between solar eruptions, radio noise from the Sun, magnetic storms and such radio fade-outs.

(iii) *Meteors*

For some years after the work of Skellett (1935) and Schafer and Goodall (1932 a, b) the only attempts to associate the increased E-region ionization with meteors appear to have been made in India. During the 1933 Leonids, Mitra, Syam and Ghose (1934) measured the electron density in the E region. Their results are given in figure 5.

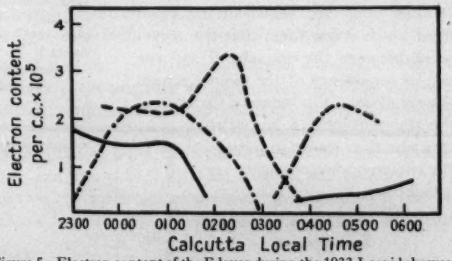


Figure 5. Electron content of the E layer during the 1933 Leonid shower.
– – – – 13 Nov. —·—·— 14 Nov. ——— 17 Nov.
(*From* Mitra, Syam & Ghose, *Nature, Lond.*, 1934, **133**, 533.)

On the night of 13–14 November the electron densities reached values of $3{\cdot}3 \times 10^5$ and $2{\cdot}2 \times 10^5$ per cm³ respectively. These workers mention that such high midnight densities are remarkable and were never recorded by them in their fortnightly observations during the 1932–33 Polar Year. These days were clear of magnetic or solar disturbances according to the records of nearby observatories.

Similar measurements were made by Bhar (1937 a, b) during the 1936 Leonids. His measurements again indicated a definite correlation of increased ionization during the night with the maximum of the shower. Bhar also looked for effects in the F region, with negative results.

Pierce (1938) made calculations which showed that the ionization due to meteors and the subsequent diffusion of the electrons could maintain the

E-region ionization at its daytime value during the night. However, his estimates are based on the theory of Maris (1929) which is now believed to be in error (Herlofson 1948), and although he gives experimental data relating to the transient abnormalities (see § 5), no experiments appear to have been made with regard to the general level of E-region ionization. In more recent work during the 1946 Giacobinid meteor shower Pierce (1947) shows records indicating that these meteors gave rise to a reflecting stratum which lasted for some hours on a frequency of 3·5 Mc/s.

Conclusive evidence that part of the abnormal-E-region ionization is associated with meteors has recently been produced by Appleton and Naismith (1947). The rate of occurrence of the transient echoes on a frequency of 27 Mc/s. was measured and compared with the occurrence of sporadic-E ionization. The results during the intense and short-lived Giacobinid shower of October 1946 are shown in figure 6, from which it is evident that the correlation between the sporadic E and the rate of occurrence of the meteor echoes is very close. The seasonal variation of the transient echoes and of the sporadic E is shown in figure 7, the correspondence is again very satisfactory.

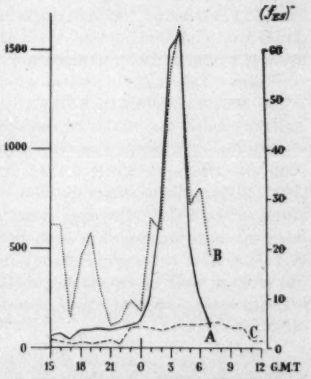


Figure 6. Showing variations of (A) ionization burst frequency, and (B) sporadic E-layer ionization during the night of the Giacobinid meteor shower. (A corresponding curve of ionization burst frequency (C) for a normal night is also shown.)

(*From* Appleton and Naismith, *Proc. Phys. Soc.*, 1947, **59**, 461.)

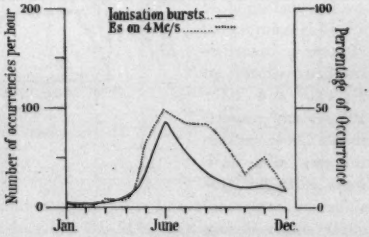


Figure 7. Seasonal variation of frequency of ionization bursts and of frequency of occurrence of sporadic E-layer ionization (4 Mc/s.) in daytime. (Average 1944–6.)

(*From* Appleton and Naismith, *Proc. Phys. Soc.*, 1947, **59**, 461.)

(iv) *The Structure of Abnormal E Ionization*

In § 3 it was mentioned that Eckersley (1929) considered that the transient scattering into the skip zone was due to ionic clouds below 125 km. A similar suggestion that the abnormal-E ionization might also be the result of clouds of extra electron concentration appears in the work of Ratcliffe and Pawsey (1933) who measured the re-

flected signals from the E region over a distance of 200 km. on wavelengths in the range 200–500 m. They concluded that scattering centres were present with an effective scattering radius of about 200 km. Maeda, Konomi and Isagawa (1935) found the abnormal effect to extend over a diameter of about 100 km. Kirby and Judson (1935 a) and Gilliland (1935) considered that the abnormalities represented a case of partial reflection at a sharp boundary, but in 1938 Best, Farmer and Ratcliffe discussed the subject in detail and concluded that under the abnormal conditions the E region contains clouds with electron densities in excess of the surrounding region and that the abnormal ionization is not in stratified form. These ideas were further developed by Appleton, Naismith and Ingram (1939) and Appleton and Naismith (1939, 1940); they concluded that the abnormal effects were due to an aggregate of such clouds, embedded in the E region, with a maximum ionization 7 km. lower than the maximum of the normal E-region ionization. The existence of such scattering clouds had been independently suggested from the work on the transient scattering centres.

(v) *Discussion of the Causes of Abnormal E*

The experimental evidence shows that there are two main causes of the long-duration abnormal-E ionization :

(*a*) Increases of ionization associated with magnetic activity due to corpuscles from the Sun.

(*b*) Increases of ionization due to meteoric ionization. Such ionization may be detected as transient responses under suitable conditions, but on wavelengths near the critical frequency, the aggregation of such effects may give rise to long-duration abnormal ionization.

Due to the deflection of the solar particles in the Earth's magnetic field, (*a*) would be expected to occur predominantly in high latitudes. The diversity of results obtained by different workers can then be readily explained, in particular, the clear correlation with meteoric activity found by the Indian workers, and the lack of correlation of abnormal E with magnetic activity found by many workers in low latitudes (e.g. Bhar and Syam 1937, Maeda, Konomi and Isagawa 1935). These views also account for the variations of abnormal ionization with latitude (e.g. Appleton, Naismith and Ingram (1937), Berkner and Wells (1937), Whatman and Hamilton (1938)).

§ 5. THE TRANSIENT ABNORMALITIES

In § 3 the early work on the short-lived transient abnormalities in the E region was described. They were associated with ionic clouds below 125 km. by Eckersley (1929) and were shown to be caused by meteoric ionization by the work of Schafer and Goodall (1932 a, b) and Skellett (1935). This conclusion was by no means unanimously accepted, however, and no further direct association of transient echoes with meteors appears to have been made until after 1940. In this section the information obtained about these transient scattering centres in the intervening period will be discussed.

The transient echoes were observed during the Polar Year Expedition by

Appleton, Naismith and Ingram (1937) who referred to Skellett's work on meteors as suggesting a possible explanation. Eckersley in 1932 extended his previous work (1929) on scattering into the skip zone and gave a theory of the scattering of radio waves from clouds of ions in the E region (see part II). Beyond pointing out that the ionization of the E region must be irregular in time and space, Eckersley did not further discuss the origin of the scattering centres. In 1937 he obtained photographic registration of the echoes from the transient clouds, using a pulsed transmitter on 9·1 Mc/s. and found that the short-lived echoes lasting for a second or so were almost continuous day and night. Among other suggestions, Eckersley referred to the fact that these echoes might be due to " small meteorites ". Skellett (1938) emphasized that Eckersley's results strongly suggested a meteoric origin. Watson-Watt, Wilkins and Bowen (1937) obtained records of these bursts of ionization between 80 and 100 km. on 6 Mc/s. and calculated the equivalent ionic densities in them to be of the order of 10^{12} ions per cm^3 lasting for several seconds.

The first detailed measurements of the transient echoes were made in 1937 by Appleton and Piddington (1938) on 8·8 Mc/s. The height distribution which these workers obtained is reproduced in figure 8. These authors considered that the echoes could be grouped into two series, one group consisting of members lasting a fraction of a second up to one or two seconds and the other of members lasting several seconds up to some minutes. From the amplitude of the returned echo it was estimated that the scattering centres consisted of clouds of 10^{16} electrons concentrated in a region of linear dimensions small compared with the wavelength. It was suggested that these results implied the entry into the atmosphere, by day and night, of an agency producing bursts of ionization of considerable intensity.

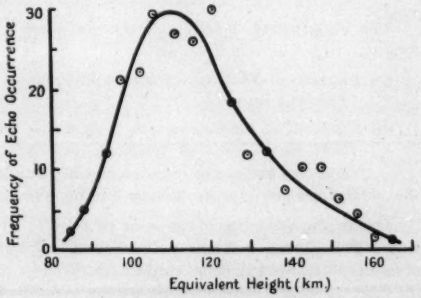


Figure 8. Height distribution of transient echoes as measured by Appleton and Piddington in 1937.
(*From* Appleton and Piddington, *Proc. Roy. Soc.* A., 1938, **164**, 467.)

In 1940 Eckersley made an extensive investigation of the echoes from an analysis of the scattering of 7 to 18 Mc/s. radio signals into the skip zone (see § 2). His results on the height distribution, diurnal variation, and the distribution of durations are reproduced in figures 9, 10, and 11.

The height distribution shows good agreement with that obtained by Appleton and Piddington (figure 8). Eckersley did not discuss the origin of these scattering clouds, but reached the conclusion that there might be a continuous gradation between these transient echoes and the longer duration

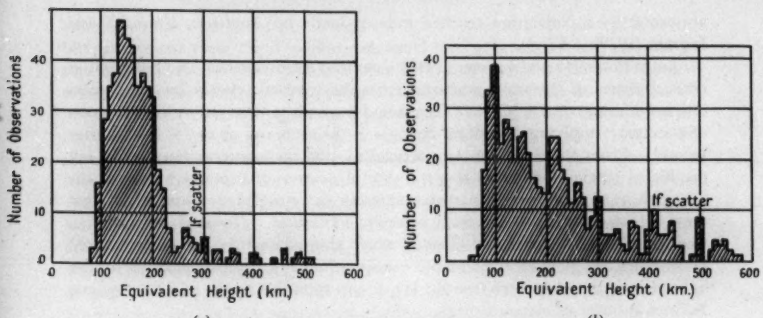


Figure 9. Height distribution of transient echoes as measured by Eckersley.
(*a*) On 39·5 m. from visual observations, Aug. to Nov. 1936.
(*b*) On 32·3 m. from daytime photographic records, April to Aug. 1936.
(*From* Eckersley, *J. Instn. Elect. Engrs.*, 1940, **86**, 548.)

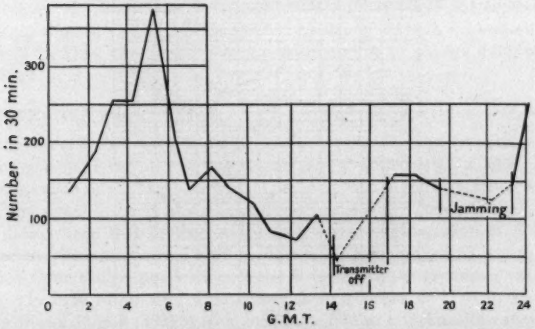


Figure 10. Ongar, 7·59 Mc/s. (39·5 m.): short-distance scattered count during 24 hours run, 15–16 October, 1938.
(*From* Eckersley, *J. Instn. Elect. Engrs.*, 1940, **86**, 548.)

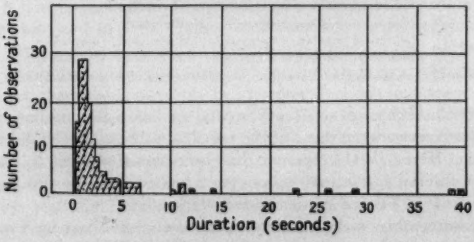


Figure 11. Ongar, 7·59 Mc/s. (39·5 m.): frequency/duration graph of the sporadic short-distance scattered echoes, Aug. to Nov. 1936.
(*From* Eckersley, *J. Instn. Elect. Engrs.*, 1940, **86**, 548.)

abnormal E—a conclusion reached independently by Appleton, Naismith and Ingram (1939) (§ 4 (iv)).

Later Eckersley and Farmer (1945) measured the directional and polarization characteristics of the wave scattered from the transient clouds by using pairs of spaced frames on 7·6 Mc/s. They found large phase changes within a fraction of a second—implying directional changes of the source of up to 15° in a quarter second. From these results they concluded that the scatter echoes could not be due to single clouds but to a number of centres distributed widely in the E region or to a shower of particles entering the atmosphere—and hence that their origin could not arise from meteoric ionization. However, the original work of Skellett, Schafer and Goodall which showed that the echoes are caused by meteors, has now received ample confirmation, and it seems possible, in the light of contemporary work (see Pt. II), to interpret the results of Eckersley and Farmer in terms of meteoric origin.

In Norway, Harang (1942) investigated the transient echoes on 6 to 8 Mc/s. He found a height distribution similar to those given by Appleton and Piddington (figure 8) and Eckersley (figure 9). Harang found durations from 0·5 to 2 or

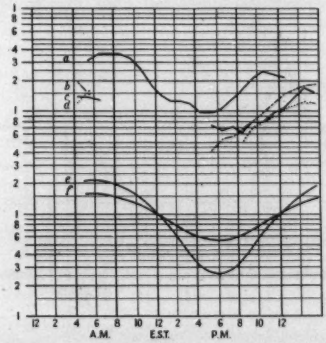


Figure 12. Diurnal variation of the numbers of bursts and meteors.
(*a*) Average of monthly median bursts/hr., Feb. 1943–Jan. 1944.
(*b*) Observed number of meteors/hr. (Schmidt).
(*c*) Observed number of meteors/hr. (Coulvier-Gravier).
(*d*) Observed number of meteors/hr. (Hoffmeister).
(*e*) Theoretical number of meteors/hr. (Hoffmeister) (parabolic).
(*f*) Theoretical number of meteors/hr. (Hoffmeister) (hyperbolic).
(*From Federal Communications Commission, Engineering Department : Docket* 6651, 1944.)

3 seconds but no echoes of one-tenth second or less. Beyond concluding that clouds of electrons were the cause of the transient echoes, Harang did not discuss their origin. Pierce (1941) repeated the observations of Skellett, Schafer and Goodall by observing transient echoes on a pulsed 3 Mc/s. equipment when meteors of the 1940 Leonid shower passed overhead.

The measurements on the transient signals described above were made either directly using pulse technique, or by observations of scattering into the skip

zone when the wave frequency is high enough to penetrate the normal E region but is reflected from the F region. A third type of phenomenon can also be caused by the occurrence of transient clouds of dense ionization in the E region—the scattering of radio signals to great distances in cases where the frequency is such that both E and F regions are penetrated and the signal is not normally received above the range of the ground ray. Taylor and Young (1928, 1929) reported abnormal scattering effects whereby high-frequency radio signals were received at a distant station after traversing a very long path. In 1931, Beverage, Peterson and Hansell found occasional signals at very great distances on frequencies as high as 40 Mc/s. In 1933 Jones observed bursts of signals up to 200 miles from transmitters working on frequencies in the range 36 to 100 Mc/s., with occasional maxima of great intensity. In 1938, Pierce discussed this type of phenomenon and pointed out that such bursts could be caused by meteor ionization. In 1942–44 workers of the Federal Communications Commission Engineering Department, using transmitters in the frequency range 42 to 84 Mc/s. and observing at distances of 100 to 340 miles, investigated these bursts in detail; visual correlation of the bursts with meteors was obtained on several occasions during August and November 1944. The diurnal variation of numbers of bursts (figure 12) shows good agreement with the theoretical and experimental data on the diurnal variation in numbers of meteors. These workers observed that on 71·75 Mc/s. the bursts were less frequent and of shorter duration than on 44 Mc/s.

§6. ECHOES FROM HEIGHTS GREATER THAN THE E REGION

Many of the earlier workers on the transient echoes looked for similar effects originating in the F region with negative results, and Bhar (1937 a, b) concluded that meteors had no measurable effect on the ionization in the F region. Ratcliffe and White (1933 b) gave an explanation of the echoes from effective heights much greater than that of the F region in terms of penetration of E region, followed by reflection from F region, reflection upwards from E and a second reflection from F then penetration of the E region to the ground—" M-type " echoes. In 1934 Mimno reported echoes from a height of 1100 to 1800 km., and Kirby and Judson (1935 a) from heights of 600 to 1100 km. (this they called the G layer). There is no evidence that these echoes were obtained other than by the mechanism suggested by Ratcliffe and White.

In 1934 Kirby, Berkner and Stuart found scattered echoes from heights of up to 1500 km. and in 1938 Harang and Stoffregen from heights of 850 to 1600 km. Appleton and Naismith (1939) then pointed out that during the Polar Year Expedition they had also obtained echoes from an apparent height of 1125 to 1425 km., but that these could only be interpreted as due to sideways scattering from the transient clouds in the E region. Eckersley (1939, 1940) observed similar effects and brought forward evidence that such echoes were due to scattering by the transient clouds in the E region, the scattered signal being returned to ground after reflection at the F layer. Harang (1942) considered that his measured heights were real, but Eckersley, Millington and Cox (1944) produced further evidence that all such effects were due to the transient scattering clouds in the E region.

It thus seems that the abnormal ionization in the E region can give a satisfactory explanation of most of the observations of echoes from apparently great heights. Wells, Watts and George (1946) claim to have detected rapidly moving ionospheric clouds moving in from 800 to 900 km. to the F region and occasionally moving out again. They attributed these clouds to an inflow of corpuscles from the Sun, but there is nothing in the published information which excludes this effect from the category considered by Eckersley above.

Recently Lovell, Clegg and Ellyett (1947) detected echoes from a luminescent auroral cloud at a height of 480 km. which persisted for 30 minutes on a frequency of 46 Mc/s.

PART II. CONTEMPORARY STUDIES OF THE TRANSIENT METEOR ECHOES

The development of radio transmitters and receivers during the war, working on frequencies considerably in excess of the critical E and F layer frequencies has given a great impetus to the study of the transient meteoric echoes. The work of Hey and Stewart (1946, 1947) using frequencies of the order of 60 Mc/s. showed, beyond any reasonable doubt, that at least on these higher frequencies the transient echoes were associated with meteors entering the Earth's atmosphere. Consequently during the last two years, considerable progress has been made in applying the technique to the study of the astronomy of meteors and to the physics of the high atmosphere.

§ 7. APPLICATION TO THE STUDY OF THE ASTRONOMY OF METEORS

(i) *The Activity of Meteors*

Hitherto the information on the activity of meteors has depended on darkness and a clear sky. The radio technique removes both these restrictions and enables an uninterrupted study to be made of the rate of meteor activity. It is now known that the number of meteor echoes observed by radio depends very critically on the wavelength (Lovell *et al.*, unpublished work). Below 6 m. wavelength the echoes are very few except when a meteor shower occurs and there is a high correlation between the transient echoes and meteors which can be observed visually (Hey and Stewart 1947, Lovell, Banwell and Clegg 1947, Prentice, Lovell and Banwell 1947). Ferrell (1946) indicates that measurements on 105 Mc/s. have been made in America, but no details are available. On wavelengths of 8 m. and above, with apparatus of equivalent sensitivity these " visual type " echoes are generally submerged in an extremely high background rate, which shows a marked seasonal and diurnal variation. Eckersley (1940), Appleton and Naismith (1947) and Eastwood and Mercer (1948) have studied these variations in detail, and there is general agreement that the echo rate shows a maximum between midnight and 0600 U.T. and a minimum during the afternoon, and that the general level of activity is higher in the autumn than in the spring. These variations can be explained if the Earth is moving through a general distribution of meteoric matter in space. There is no close correlation between visible meteor showers and this high background rate exhibiting the diurnal and seasonal variation.

Figure 13 illustrates these effects. It shows the rate of transient echoes observed simultaneously with 4 m. and 8 m. apparatus with similar aerial systems

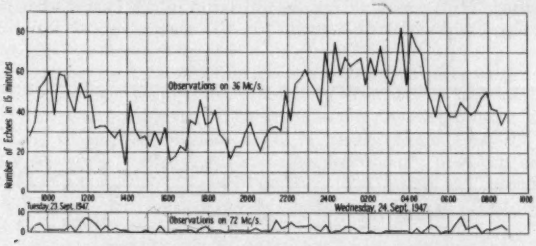


Figure 13. Simultaneous observation of transient echo rates on 72 and 36 Mc/s. Tuesday 23 Sept. to Wednesday 24 Sept. 1947.

during a 24-hour watch from 0900 U.T. on 23 September to 0900 U.T. on 24 September 1947. There are two important points about these observations :

(*a*) There is conclusive experimental evidence from the work of Hey and Stewart (1946, 1947), Prentice, Lovell and Banwell (1947), and Lovell, Banwell and Clegg (1947) that the echoes shown in the lower curve are associated with meteors, of the type which occur in the recognized visual showers.

(*b*) In this experiment all the echoes observed on 4 m. were also observed on 8 m., but in addition there was an extremely high background rate on 8 m. with a diurnal variation. There is, as yet, no *direct* experimental evidence that this high echo rate is due to meteors ; it is generally accepted that it is so, (i) because the echoes produced are indistinguishable from those produced by meteors for which visual correlation can be obtained, and (ii) because the seasonal and diurnal variation is that to be expected from a general distribution of meteoric matter through which the Earth may be passing.

The exact nature of the second type of " meteor echoes " remains to be determined. If, as seems highly probable (Appleton and Naismith 1947), they are due to a uniform distribution of meteoric dust, too minute to produce visible meteors, then the question whether the dust is confined to the solar system or exists in interstellar space is a matter of considerable astronomical interest.

(ii) *The Study of the Major Meteor Shower Rates*

Apart from exceptional cases of very high shower rates the background echo rate referred to above obscures the influence of the ordinary meteor showers on radio apparatus working on wavelengths of 8 m. and above. In the 4 m. region the background rate is very low and due only to activity from minor radiants and occasional single meteors. Hence the major meteor showers can be studied in detail. The first measurements on the major showers were made

by Hey and Stewart (1946, 1947) on a wavelength of 4 to 5 m. Figure 14 shows the mean hourly echo rate obtained with their apparatus during January and April 1946.

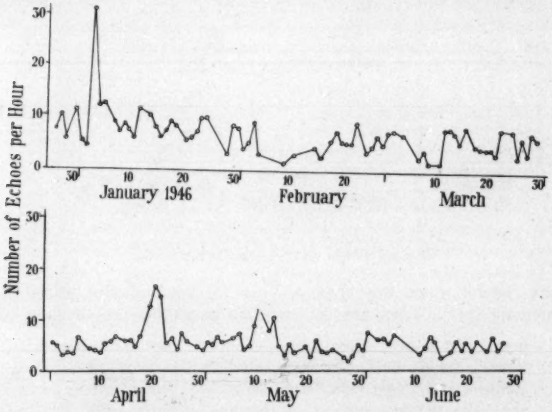


Figure 14. Mean hourly rate of occurrence of echoes. Vertical-beam station, Richmond Park. (*From* Hey and Stewart, *Proc. Phys. Soc.*, 1947, **59**, 858.)

The peak in hourly rate in January corresponds to the Quadrantid meteor shower and that in April to the Lyrids, both agreeing in time with the visible

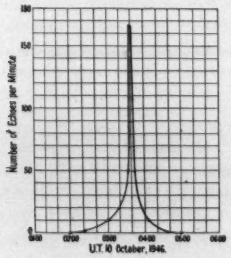


Figure 15. The rate of transient echoes from the Giacobinid Shower, 10 October 1946, measured from ciné film photographs of the cathode ray tube.

Half-wave dipoles as aerial system on 72 Mc/s.

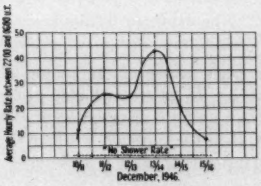


Figure 16. Average nightly (hourly) rate of Geminids, Dec. 1946. Observed on 72 Mc/s.

occurrence of these showers. Similar curves of the activity during the summer of 1946 including the Perseid shower are given by Prentice, Lovell and Banwell (1947).

The remarkable Giacobinid meteor shower on 10 October 1946 was studied by Appleton and Naismith (1947), Hey, Parsons and Stewart (1947), and Lovell, Banwell and Clegg (1947) in Great Britain, and by Pierce (1947) in America. The echo rate during the night of 9–10 October 1946 obtained by Appleton and Naismith (1947) is shown in figure 6.

The frequency used was 27 Mc/s. and this is an exceptional case of a visible type of meteor shower giving a very pronounced effect over the high background rate found on these frequencies. The corresponding rate curve found by Lovell, Banwell and Clegg (1947) working on 72 Mc/s. is shown in figure 15. The normal " no-shower " rate on this apparatus was about 2 per hour, and this condition applied until 00^h U.T. on 10 October. But by 0340 U.T. the echo rate had risen to the enormous peak of 168 per minute and by 0600 U.T. had decreased again to the normal " no-shower " rate of 2 per hour.

Since October 1946 the activity of the major meteor showers has been investigated systematically by Lovell and co-workers on a frequency of 72 Mc/s. The case of the Giacobinid shower is, of course, unique both in its activity and in the shortness of its duration. The more normal behaviour of the major showers may be illustrated by the rate curve of the Geminid shower shown in figure 16 for December 1946.

(iii) *The Determination of Meteor Radiants*

Pierce (1938) first drew attention to the influence of the orientation of the meteor trail with respect to the observing apparatus, and subsequent experimental work (Hey and Stewart 1946, 1947, Lovell, Banwell and Clegg 1947) has shown that meteor trails do, in fact, show a critical aspect effect, and that the main radio echo is obtained only when the aerial beam is directed at right angles to the trail. This property of meteor trails has been used to determine the radiants of showers by observations of the radio echoes. Two methods have so far been devised :

(a) *Method of Hey and Stewart.* The first determination of meteor radiants by radio methods was made by Hey and Stewart (1946, 1947). These workers used three separate stations, operatiug on 73 Mc/s., situated as shown in figure 17.

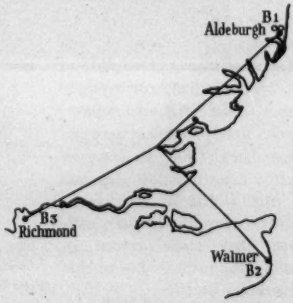


Figure 17. Equipments in operation, June and July 1945.
(*From* Hey and Stewart, *Proc. Phys. Soc.*, 1947, **59**, 858.)

The direction of the aerials was adjusted so that the beams intersected at a point about 100 km. in height equidistant from each station. Since a meteor trail will give a radio echo only when it passes at right angles to the axis of the beam, it is to be expected that as a radiant moves across the sky it will not produce echoes simultaneously at the three stations, but in succession as the meteor trails become oriented at right angles to the respective aerial beams. The diurnal variations of the

mean hourly rates obtained by Hey and Stewart on the stations B1, B2, B3 (figure 17) between 26 July–1 August 1945 are shown in figure 18.

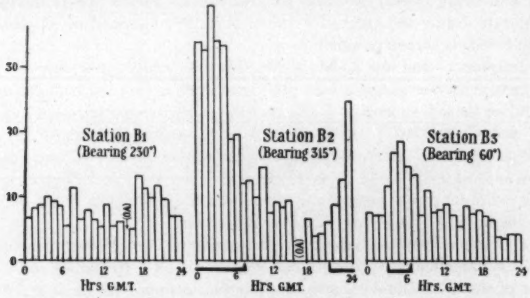


Figure 18. Diurnal variations of mean hourly rate of echoes, 26 July – 1 August 1945. Times at which the radiant R is favourable are indicated by heavy lines
Ordinates = relative number. Abscissae = time G.M.T. O.A. = out of action.
(*From* Hey and Stewart, *Proc. Phys. Soc.*, 1947, **59**, 858.)

The coverages of possible radiant points required to produce the peaks in echo rate for B2 at 0230 and B3 at 0430 are shown in figure 19. The centre of overlap R may be taken to be the radiant position and gives a radiant point at R.A. 345°, Decl. −10°. The station B1 shows no marked peak. This is to be expected since a radiant in this position never presents an aspect favourable to B1. This radiant corresponds well with the δ-Aquarid radiant, known to be prominent between these dates. The accuracy of such a determination depends on the width of the aerial beams. Hey and Stewart estimate that their errors in placing the radiant due to this reason may be 10°.

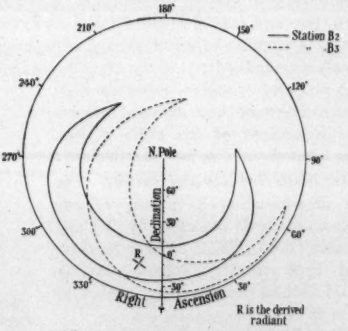


Figure 19. Coverage of possible radiant positions for main peaks in hourly rate of stations B 2 and B 3.
(*From* Hey and Stewart, *Proc. Phys. Soc.*, 1947, **59**, 858.)

(b) *Method of Clegg.* Another method of determining meteor radiants using a movable narrow beam aerial at a single observing station has been developed by Clegg (1948). In order to determine the right ascension of a radiant an aerial beam is directed due east. No echoes will be observed until the meteor trails pass through the beam at right angles to the normal from the observing station, that is, until the radiant is due south. By considering the geometry of the intersection of the aerial beam with the 100-km. zone in which the meteors

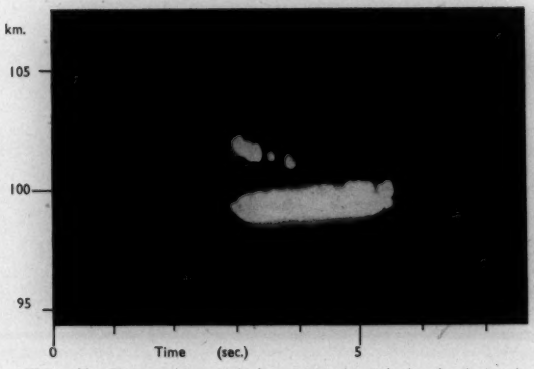


Figure 21. Photographic record of a transient ionospheric echo during the Giacobinid meteor shower 1946.
(*From* Hey and Stewart, *Proc. Phys. Soc.*, 1947, **59**, 858.)

To face page 432

ionize, it can be shown that these echoes will appear at long range and that as the radiant moves further west the range of the echoes will decrease steadily. Finally, as the radiant moves out of the beam the echoes will close in to short range and disappear. The time at which the long range echoes first appear will correspond with the time of transit, from which the right ascension of the radiant can be calculated.

In order to determine the declination of the radiant, the aerial beam is now rotated through some angle θ, towards the south, so that the beam is again in advance of the radiant. As the radiant moves into the position at right angles to this new direction the echoes will again appear at long range and their ranges will decrease as the radiant moves still further west. From the time interval between the first appearance of the long range echoes in each of the two positions the time taken for the radiant to move through an angle θ after transit is known, and hence the declination of the radiant can be calculated. Figure 20 indicates

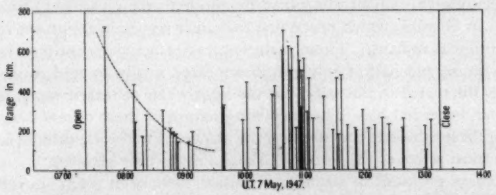


Figure 20. Range/time plots of echoes observed with narrow beam aerial directed eastwards on 7 May 1947, as an illustration of Clegg's method for determining radiant positions. Plots between $07^h\ 30^m$ and $09^h\ 30^m$ show transit of η-Aquarids through the beam (– – – – theoretical distribution). Subsequent echoes represent a new daylight shower with a marked transit beginning at $10^h\ 35^m$ U.T. (the Piscids).

the kind of results obtained by this method, and shows the ranges of individual echoes observed on 7 May 1947 with the beam directed due east. The first stream was the η-Aquarids, with transit time 0740 U.T. When this stream had passed through the beam a second radiant appeared at 1040 U.T., finally disappearing at about 1230 U.T. Further readings on a second azimuth enabled the declination to be determined and showed that this second shower was from a radiant in Pisces (see § 7 (iv)).

The accuracy of the method depends on the horizontal beam width of the aerial, the echo rate, and the declination of the radiant. For a given beam width the technique is capable of considerably greater accuracy than the method of Hey and Stewart since it makes use of the measured ranges of the echoes. With the apparatus used in the investigation it was estimated that the average radiant giving echoes at a rate of about 30 per hour could be placed within an ellipse of semi-major axis 3° and semi-minor axis 1°, the minor axis being parallel to the celestial equator.

(iv) *The Discovery of Daylight Meteor Showers*

The methods described above enable meteor showers to be studied in daylight and the preliminary results obtained show some very surprising features. By the method described in §7 (iii, a) Hey and Stewart (1947) found a daylight

shower active between 6–13 June 1945 at R.A. 58°, Decl. +5°, and the measurements of Prentice, Lovell and Banwell (1947) during the summer of 1946 indicated the presence of considerable activity during the daytime. The systematic study of the major showers by Lovell and his co-workers commencing in October 1946 showed that throughout the autumn and winter the general predictions of shower activity made from visual observations were substantially correct. But in the investigation of the η-Aquarid radiant which began on 1 May 1947, it was found that the meteor stream visible to the astronomers was merely the beginning of an extremely active belt stretching towards the Sun. The radiant positions were determined by Clegg's method and the main radiant point of the stream was found to lie in Pisces; since transit occurred between 1000 and 1100 U.T. no knowledge of the shower can have been obtained by visual observation.

The Aquarius end of this belt disappeared in mid-May as expected, but the Piscid radiant continued with undiminished and extraordinary intensity. On 18 June the original radiant was joined by another very vigorous radiant further to the east in Gemini, which proceeded to transit regularly about one hour later than the original radiant. These main centres of activity continued at a high rate until the second half of July and disappeared finally by mid-August. The main knots fluctuated in intensity, but the hourly rate at transit rarely fell below 20 and often exceeded 80. The surprising nature of these events is evident by comparing these rates and the extent of the shower with the Geminids (figure 16), one of the most active of the recognized major night-time showers.

Preliminary positions of the Piscid radiant have been given (Lovell 1947 a) as follows:

May	7·47	R.A. 20°	N. Decl. 20° to 25°,
June	22·47	R.A. 70°	N. Decl. 20° to 25°,

with minor radiants 15° W. and 10° E. Occasionally the whole area between these extremes was active.

The two most surprising features of this daylight stream are its duration—almost one-quarter of the year—and the constancy in distance from the Sun of the main radiant points. As yet there is no satisfactory astronomical explanation of this phenomenon.

(v) *The Determination of Meteor Velocities* *

During the 1946 Giacobinid shower Hey, Parsons and Stewart (1947) succeeded in making the first determination of the velocities of the meteors from observations of the radio echoes. Instead of the more usual presentation in which the signals are displayed as deflections of a linear range trace on a cathode-ray tube, the received echoes modulated the brightness of a linear-range trace. This was photographed on a 35-mm. film moving at right angles to the trace with a uniform speed of 2·4 mm. per sec. By this technique a continuous photographic record was obtained of the range of the echoes and of their time of occurrence and

* *Note added in proof.*—Ellyett and Davies, working at the Jodrell Bank Experimental Station, have recently succeeded in measuring meteor velocities by a new method of more general applicability than that described here. The method consists in measuring the fluctuations in the received power as the meteor crosses the perpendicular from the station to the direction of the trail. The maxima and minima in the fluctuations correspond to the formation of successive Fresnel zones from which the velocities have been determined (C. D. Ellyett and J. G. Davies, *Nature, Lond.*, 1948, **161**, 596).

duration. Several of these photographic records showed a remarkable faint fast-moving echo prior to the formation of the main echo. An example is shown in figure 21 in which a faint streak can be observed beginning at coordinates of approximately (103 km., 1·75 sec.) and moving in to the main echo extending from approximately 2·75 to 5 seconds at a constant range of just under 100 km.

Hey, Parsons and Stewart attribute this faint streak to energy scattered from the ionization at the head of the approaching meteor and the main intense echo to the ordinary return of energy when the meteor trail approaches the point of intersection of the normal from the observing station to the ionized column (see §7 (iii) above). The faint track thus gives the relation between the slant range and time for the approaching meteor. If the meteor has a velocity V in a straight line, then the range R at any instant T is given by

$$R^2 = R_0^2 + V^2(T - T_0)^2,$$

where R_0 and T_0 represent range and time at which the range is a minimum. V was determined by taking three observed values of R and T and eliminating R_0 and T_0. Hey, Parsons and Stewart computed the velocity of 22 tracks of the Giacobinid meteors in this manner and obtained a mean value of $22{\cdot}9 \pm 1{\cdot}3$ km. per sec. in excellent agreement with the theoretical geocentric velocity of 23·7 km. per sec.

§8. APPLICATIONS TO PHYSICS AND THE STUDY OF THE UPPER ATMOSPHERE

(i) *Intensity of the Scattered Radio Waves*

(a) *Eckersley's calculations.* The experiments of Eckersley (1932, 1940) on the transient echoes were made on wavelengths exceeding 15 m., at which wavelengths

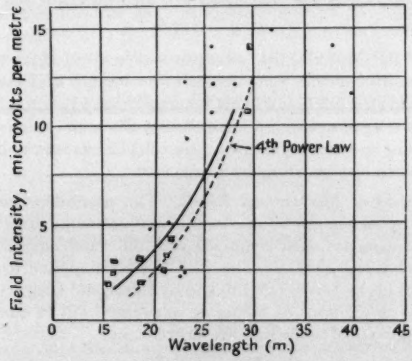


Figure 22. Intensities of scattering as a function of wavelength, from observations on commercial stations by reducing the values to a standard 5 kw. output with 4th-power law for comparison. (*From* Eckersley, *J. Instn. Elect. Engrs.*, 1940, **86**, 548.)

the transient echoes were made on wavelengths exceeding 15 m., at which wavelengths the background rate showing the diurnal and seasonal variation is predominant over the effects of the shower meteors (see §7 (i)). He considered that the

echoes were due to ionic clouds existing for more than 0·02 sec. but less than 1·0 sec. and treated the scattering of radio waves from such clouds as analogous to the scattering of α-particles by heavy atoms. On this theory the back-scattered intensity should be proportional to λ^4. In figure 22 Eckersley's experimental values (1940) for the intensity of scattering over the wavelength range 15 to 30 m. are compared with the predicted fourth power law.

(b) *Pierce's calculations.* Pierce (1938) calculated the scattered energy from a meteor trail compared with that scattered from an ionized layer by assuming that the electrons in the trail formed a long column with a diameter *large* compared with the wavelength.

In figure 23 AB is the equivalent reflecting plane of an ionized layer and the circle with centre C, the equivalent reflecting cylinder containing a meteor track along its axis. The energy radiated from O and reflected from AB to O to fall upon a small area at O is

$$E_p = k/h^2.$$

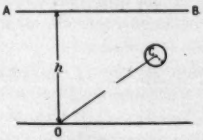


Figure 23.

For the cylinder at C, if the density perpendicular to the axis of the cylinder is proportional to r/d and uniform along the axis, the reflected energy is

$$E_c = (k/d^2)(r/d) = kr/d^3.$$

Thus
$$E_c/E_p = rh^2/d^3.$$

Pierce then takes as an example the case of reflection for a meteor trail at $d = 100$ km. with $r = 0{\cdot}25$ km., compared with that for the F region at 250 km., which gives
$$E_c/E_p = 0{\cdot}0156.$$

The ratio of the respective field strengths is then $\sqrt{(E_c/E_p)}$ or 12·5/100.

Pierce mentions that this ratio compares favourably with the experimental values of the field strength found in his " bursts " (see § 5), but he gives no other experimental data against which it is possible to assess the validity of this treatment. According to Herlofson (1948) these calculations lead to densities in the meteor trails which may be wrong by a large factor.

(c) *Calculations of Blackett and Lovell.* The possibility that some of the transient radio echoes might be due to coherent scattering from electrons in a long column of diameter *small* compared with the wavelength λ was suggested by Blackett and Lovell (1941). This idea has been applied to the particular case of meteor trails by Lovell (1947 b) and by Lovell and Clegg (1948). In this case the number of electrons N, scattering coherently, will be those contained in a column of length equal to the first Fresnel zone, i.e.

$$N = \alpha\sqrt{(\lambda R/2)},$$

where α = number of electrons per centimetre of the path formed by the meteor. By introducing the parameters of the radio apparatus it can then be shown that

$$\alpha = \sqrt{24\pi}\left(\frac{mc^2}{e^2}\right)\frac{1}{G}\sqrt{\frac{\epsilon R^3}{P\lambda^3}} \qquad \ldots\ldots(1)$$

where G is the power gain of the receiving and transmitting aerials over a half-wave dipole, ϵ is the scattered power delivered to the receiver, R is the distance of the trail, P the peak transmitter power and mc^2/e^2 the inverse of the classical electron radius.

The predicted relation between ϵ and λ has been tested experimentally (Lovell 1947 b, Lovell and Clegg 1948) by making simultaneous observations of radio echoes from the same meteor trail on different wavelengths. Within the limits of error of the experiment the relation found agreed with the above predictions over the range of wavelength used, viz. 1·4 m. to 8 m. The measured values of α for visible meteor trails using this formula are also in good agreement with theory (Herlofson 1948).

(d) *Discussion of the scattering formulae.* As yet there is insufficient experimental data to decide which of the above treatments is correct. Lovell's treatment yields a formula which agrees with experiment as regards the variation of ϵ with λ over the range of wavelengths so far tested, 1·4 to 8 m., and gives results for the electron density in good agreement with theory (Herlofson 1948). It also predicts correctly the variation in numbers of echoes recorded as λ changes over the range 1·4 to 6 m. It fails completely, however, to account for the sudden increase in echo rate on 8 m. as compared with 6 m. described in §7 (i) above. Eckersley's formula agrees with experiment over the range 15 to 30 m., but does not apply over the range of wavelengths below 8 m., where the shower meteors are predominant. Both Pierce's and Eckersley's formulae give values of electron density which are in disagreement with contemporary theoretical estimates (Herlofson 1948).

It seems possible that the large background rate on the longer wavelengths may be due to meteoric dust producing electron columns with different properties from those produced by the shower meteors. In fact it is difficult to see how the properties of the two types of ionization can be the same, since, as mentioned in §7 (i) above, the " invisible background meteors " give the same kind and amplitude of response in 8 m. radio observations as those associated by experiment with the visible meteor trails.

(ii) *Electron Density in Meteor Trails*

By carrying out a combined visual and radio-echo watch it is possible to measure the amplitude of the radio echo produced by a visible meteor. This was done for the 1946 Perseid shower by Prentice, Lovell and Banwell (1947), for the 1946 Giacobinid shower by Lovell, Banwell and Clegg (1947) and for the 1947 Perseid shower (Lovell 1947 b). By using equation (1) it is possible to calculate α, the number of electrons produced per centimetre of path by the meteor, since G, P, λ are known constants of the apparatus and ϵ and R are measured for each trail. The three sets of results are in good agreement and yield values for α between 10^{11} and 10^{12} electrons per centimetre of path. Detailed associations of visual magnitude with α have not yet been made, but the results show that a meteor of about fifth magnitude, near the limit of naked-eye visibility, produces about 2×10^{10} electrons per centimetre of path. This result is in good agreement with the theoretical estimates of Herlofson (1948). It is of interest

to note that if these estimates are correct a visible meteor has a radius of approximately $\frac{1}{10}$ mm. and weighs a few milligrams.

Pierce (1947) calculated the energy brought into the atmosphere by the entire Giacobinid shower of October 1946 and obtained a rate of flow of 3 watts per km^2. On the basis of Maris's theory (1929) of meteor ionization he then calculated the rate of ion production as 42 ions per cm^2 per sec. According to Herlofson (1948) these estimates may be in error by a considerable factor.

(iii) *The Durations of the Transient Echoes*

The distribution of durations of the transient echoes have been studied by many workers (Appleton and Piddington 1938, Eckersley 1940, Hey and Stewart 1947, Prentice, Lovell and Banwell 1947) and all agree in assigning a distribution of the type obtained first by Eckersley and shown in figure 10. Recent work in which the same meteor trail has been observed simultaneously on two wavelengths (Lovell and co-workers, unpublished) has also confirmed earlier suggestions (Eckersley 1940, Federal Communications Commission 1944) that the durations of the echoes increase as the wavelength increases. Thus, although the general shape of figure 10 remains the same, the lower limit shifts to shorter durations as the frequency increases. In their measurements on 27 Mc/s. Appleton and Naismith (1947) found that the durations were longer after midnight than before and interpreted this result as indicating that stronger reflections were obtained in the second half of the night than in the first.

As yet there is no satisfactory explanation of the factors controlling the duration of these transient echoes. It is clear that the height of formation in the atmosphere is not the only determining factor. Although the heights of individual echoes have not yet been measured with precision by any workers, it is a common observation that echoes in the height range 100 ± 5 km. can have durations varying over a range of 1000 to 1. Further, the systematic study of the major showers by Lovell and co-workers has shown that the distribution in duration of echoes from the major showers do not show any marked differences. Figure 24 shows the results of the distribution of durations of the 1946 Geminid, Leonid, Orionid and Giacobinid showers and the 1947 Quadrantid shower. According to Porter (1944) the velocity of these meteors extends from 23·7 km. per sec. (Giacobinids) to 72 km. per sec. (Leonids) and visual observations (Whipple 1943, Prentice 1948) show a clear connection between the height of appearance of the meteor trail and the velocity of the meteor. Those in the lower velocity group become visible at approximately 95 km. altitude and those in the higher velocity group at 121 km. altitude. According to information given by Harang (1946) the atmospheric pressure varies by a factor of 10 over these altitudes and if recombination or attachment processes alone controlled the duration of the echoes, this pressure effect should be visible in figure 24. From the similarity in the distributions plotted in this figure it is therefore evident that the rate of recombination and attachment as determined by the height of appearance of the meteor in the atmosphere is not the only factor controlling the duration of the radio echo.

Herlofson (1948) has suggested that the diffusion of the electrons and not recombination may be the primary factor determining the duration of the radio

echoes. The direction of the meteor trail relative to the Earth's magnetic lines of force will determine whether diffusion takes place in such a manner that the electron column is broadened or lengthened. The duration of the radio echo will thus depend on whether the aerial is directed so that the echo is observed from a column which is lengthening or broadening. In the latter case interference effects will quickly reduce the scattered intensity and the echo will last for a shorter time than in the former case.

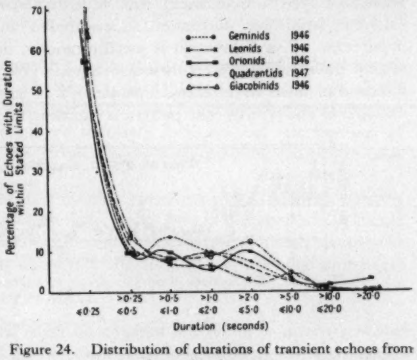


Figure 24. Distribution of durations of transient echoes from five major showers.

The existing experimental evidence is very much in favour of Herlofson's theory. (*a*) If diffusion is controlling the duration, then for given sensitivities of equipment the duration of the echo should depend on the wavelength. The simultaneous observations on two wavelengths of echoes from the same meteor trail indicate that the duration is approximately proportional to λ^2. (*b*) If the direction of the trail relative to the Earth's field is significant then the duration of the echoes from a given meteor shower should depend on the direction of the aerial beam. In the measurements made by Lovell and co-workers at the Jodrell Bank Experimental Station (53° 13′ 53″ N., 2° 18′ 11″ W.), using the directional aerial system, the average duration of the echoes as a function of azimuth of the aerial in a northerly direction has been measured for the 1946 Orionids, Leonids, and Geminids and the 1947 Piscids. In each case a marked peak in the average duration of the echoes was obtained between 0° and 20° E. The results for the Geminids and Piscids are shown in figure 25, where the average duration of the echoes is plotted as a function of the azimuth of the aerial beam.

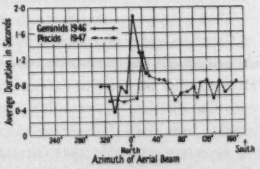


Figure 25. Average duration of echoes from 1946 Geminids and 1947 Piscids as a function of direction of observation.

(iv) *Fluctuations and Movements of the Echoes and the Possible Application to the Study of Winds*

The radio echoes obtained from meteor trails show very marked fluctuations in intensity during their life. A technique of studying this by ciné-photography for the long duration echoes and examples of the fluctuations obtained have been given by Lovell, Banwell and Clegg (1947). Subsequent more detailed

studies from ciné-photographs of several hundred meteor echoes have been made by Ellyett (unpublished), who finds that nearly all the echoes show a considerable amplitude fluctuation superimposed on the decay from the initial maximum. This fluctuation is usually random between successive one-eighth second photographs, but sometimes shows a grouping into subsidiary maxima a second or more in duration, separated by an interval of several seconds. An example of this type of echo pattern is shown in figure 26, where the echo ampli-

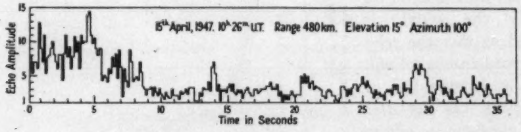


Figure 26. Fluctuations of echo amplitude with time as measured on 72 Mc/s. Echo from Piscid radiant on 15 May 1947.

tude is given in terms of the background noise level as unity and the time in seconds is taken from an arbitrary zero near the commencement of the echo. Herlofson (1948) has suggested that such variations may be connected with the distortions of the ionized track by high altitude winds.

Ellyett has also shown that a very high speed fluctuation with a period of the order of 1/10 second or less is superimposed on the slower period fluctuation mentioned above. As yet there is no explanation of this phenomenon.

Hey and Stewart (1947) found that 2% of all the echoes they observed showed a movement in range and that 15% of echoes with duration in excess of 1 sec. showed a movement. For echoes exceeding 0·5 sec. in duration these workers found that the radial velocities of such movements were mainly between 0 and -5 km. per sec. with a few extending to $+5$ and -25 km. per sec. The majority of these movements must represent a different phenomenon from the occasional detection of the ionization at the head of a travelling meteor used by Hey, Parsons and Stewart to determine velocities (§ 7 (iii)), and are probably connected with the drifting of the column of ionization under the influence of high altitude winds.

It seems possible that further investigations of the fluctuations and drifting of the echoes may yield information about the winds at high altitude.

(v) *The Distribution of Meteoric Mass*

As discussed in § 8 (ii) above, the observation of the transient echoes enables the electron density in the meteor trail to be computed. By assuming that equation (1) above was correct for the shower meteors, Lovell, Banwell and Clegg (1947) computed the distribution of electron densities in the meteors of the 1946 Giacobinid shower from observations, on a frequency of 72 Mc/s., of the echo amplitudes and ranges measured from the ciné film taken at the peak of the shower. The distribution in number of these densities was found to obey an inverse power law with an exponent of $0·9 \pm 0·2$. Further measurements on the 1946 Geminids gave a value for the exponent of $0·9 \pm 0·2$ and measurements on the 1947 Quadrantids a value of $1·0 \pm 0·2$. It therefore seems

that the distribution in electron densities amongst the shower meteors conforms to an inverse power law with an exponent close to unity over the ranges of electron densities covered by the measurements (10^9 to 10^{12} electrons per centimetre of path).

If the velocities of the meteors in a given stream are assumed to be constant it follows as a consequence of Herlofson's theory that the distribution of the masses of the meteors in the stream must be the same as the distribution of electron densities. The above measurements therefore indicate that the number of meteors of a given mass in these streams is inversely proportional to the mass of the meteors.

This result is in agreement with existing visual observations which indicate that there is the same total mass of meteoric material in each magnitude (Fletcher Watson 1941). The future determination of such mass distributions by observation of the radio echoes will be of considerable interest when applied to meteors below the visibility limit.

(vi) *The Phenomenon of "Whistling Meteors"*

In 1941 Chamanlal and Venkataraman heard weak whistles of an unusual character when they tuned in to an unmodulated carrier on a frequency of 7 Mc/s. at a distance of 10 miles from the Delhi short-wave transmitter. The whistles appeared as high-pitched notes rapidly descending in pitch and of short duration—usually from about one-fifth seconds to several seconds—and occurring at random time intervals. By visual correlation Chamanlal and Venkataraman showed that these whistles were associated with the passage of a meteor through the atmosphere. According to these workers the explanation of the whistles is that they are heterodyne notes caused by interference of the radio wave scattered from the ionization at the head of the approaching meteor with the direct ground ray. To account for the decreasing pitch of the note it was assumed that the velocity of the meteor was retarded as it advanced into the atmosphere ; the scattered wave on interfering with the ground waves then gives a whistle of gradually descending pitch.

These experiments were discussed by Khastgir (1943) who questioned the above explanation of the whistles and suggested that some of the whistles might be due to atmospherics and some due to meteors producing an "electrical impulse" when they stopped in the atmosphere. However, the association of the whistles with meteors has since received confirmation through the work of Manning, Helliwell, Villard and Evans (1946), Pierce (1947) and Appleton and Naismith (1947). The former workers listened on two equipments during the 1946 Giacobinid shower (*a*) 72 miles from a 100 kw. transmitter on 15 Mc/s. (*b*) 1 mile from a 0·7 kw. transmitter on 29 Mc/s. They observed Doppler whistles coincident with visual meteors on both receivers, and the whistles were often followed by bursts of strong signal. Pierce (1947) also observed the Doppler whistles during the Giacobinid shower and considered that the explanation of Chamanlal and Venkataraman, that the whistles were due to retardation of the meteor, to be incorrect. Pierce suggests that the Doppler beat is produced by an "ionic bow wave" and is a function of both the velocity of the meteor and the rate of diffusion of the electrons.

Appleton and Naismith (1947) also consider the explanation due to retardation of the meteor to be incorrect and show that the pitch of the whistle should fall if the wave is scattered from the head of the meteor as it approaches with constant velocity. When the meteor reaches the point of intersection of the normal from the station to the direction of travel of the meteor, the pitch should fall to zero. But this is the favourable point at which the meteor trail gives rise to the main radio echo and hence the whistles should always precede the echo if this explanation is correct. The observations of Manning, Helliwell, Villard and Evans (1946) that the bursts of strong signal followed the whistles are in favour of this explanation.

The work of Hey, Parsons and Stewart (1947) on the velocities of the meteors (§7 (v)) indicate that this explanation is almost certainly correct. These workers found no measurable retardation of the meteor during their velocity measurements, but obtained photographic records of the fast moving echo due to the ionization at the head of the approaching meteor (see figure 21).

ACKNOWLEDGMENTS

The author is indebted to Mr. J. S. Hey for supplying the information on the determination of meteor velocities prior to publication and to Mr. N. Herlofson for much helpful discussion in connection with Part II of this report. He also wishes to acknowledge his debt to his colleagues at the Jodrell Bank Experimental Station of the University of Manchester who shared in the hitherto unpublished work referred to in Part II, particularly to Mr. J. A. Clegg and Mr. C. D. Ellyett, who also helped with the preparation of Part II of the report, and to Messrs. C. J. Banwell, J. G. Davies, D. R. H. Forsyth, V. A. Hughes, and F. Moran. Finally, the author wishes to thank Prof. P. M. S. Blackett, who has given all possible aid and encouragement to the development of the Experimental Station at Jodrell Bank, and hence provided the initiative for the writing of this Report.

REFERENCES

Appleton, E. V., 1927 a, *Nature, Lond.*, **120,** 330 ; 1927 b, *Electrician*, **98,** 256 ; 1930, *Proc. Roy. Soc.*, A, **126,** 542 ; 1935, *Phys. Soc. Rep. Progr. Phys.*, **2,** 129 ; 1939, *Occ. Notes, R. Astr. Soc.*, **3,** 33.

Appleton, E. V., and Barnett, M. A. F., 1925 a, *Nature, Lond.*, **115,** 333; 1925 b, *Proc. Roy. Soc.*, A, **109,** 621.

Appleton, E. V., and Green, A. L., 1930, *Proc. Roy. Soc.*, A, **128,** 159.

Appleton, E. V., and Hey, J. S., 1946, *Phil. Mag.*, **37,** 73.

Appleton, E. V., and Naismith, R., 1932, *Proc. Roy. Soc.*, A, **137,** 36 ; 1933, *Proc. Phys. Soc.*, **45,** 389 ; 1939, *Nature, Lond.*, **143,** 243 ; 1940, *Proc. Phys. Soc.*, **52,** 402 ; 1947, *Ibid.*, **59,** 461.

Appleton, E. V., Naismith, R., and Builder, G., 1933, *Nature, Lond.*, **132,** 340.

Appleton, E. V., Naismith, R., and Ingram, L. J., 1937, *Phil. Trans. Roy. Soc.*, A, **236,** 191 ; 1939, *Proc. Phys. Soc.*, **51,** 81.

Appleton, E. V., and Piddington, J. H., 1938, *Proc. Roy. Soc.*, A, **164,** 467.

Appleton, E. V., and Ratcliffe, J. A., 1930, *Proc. Roy. Soc.*, A, **128,** 133.

Appleton, E. V., Watson-Watt, R. A., and Ratcliffe, J. A., 1933, *Proc. Roy. Soc.*, A, **141,** 697.

Berkner, L. V., and Wells, H. W., 1937, *Terr. Magn. Atmos. Elect.*, **42,** 73.

Best, J. E., Farmer, F. T., and Ratcliffe, J. A., 1938, *Proc. Roy. Soc.*, A, **164,** 96.

Beverage, H. H., Peterson, H. O., and Hansell, C. W., 1931, *Proc. Inst. Radio Engrs.*, **19,** 1313.

BHAR, J. N., 1937 a, *Nature, Lond.*, **139,** 470 ; 1937 b, *Indian J. Phys.*, **11,** 109.
BHAR, J. N., and SYAM, P., 1937, *Phil. Mag.*, **23,** 513.
BLACKETT, P. M. S., and LOVELL, A. C. B., 1941, *Proc. Roy. Soc.*, A, **177,** 183.
BREIT, G., and TUVE, M. A., 1925 a, *Terr. Magn. Atmos. Elect.*, **30,** 15; 1925 b, *Nature, Lond.*, **116,** 357 ; 1926, *Phys. Rev.*, **28,** 554.
CHAMANLAL, and VENKATARAMAN, K., 1941, *Electrotechnics*, **14,** 28.
CLEGG, J. A., 1948, in preparation.
DELLINGER, J. H., 1935, *Phys. Rev.*, **48,** 705.
EASTWOOD, E., and MERCER, K. A., 1948, *Proc. Phys. Soc.*, in press.
ECKERSLEY, T. L., 1929, *J. Instn. Elect. Engrs.*, **67,** 992 ; 1932, *Ibid.*, **71,** 405 ; 1937, *Nature, Lond.*, **140,** 846 ; 1939, *Ibid.*, **143,** 33 ; 1940, *J. Instn. Elect. Engrs.*, **86,** 548.
ECKERSLEY, T. L., and FARMER, F. T., 1945, *Proc. Roy. Soc.*, A, **184,** 196.
ECKERSLEY, T. L., MILLINGTON, G., and COX, J. W., 1944, *Nature, Lond.*, **153,** 341.
FEDERAL COMMUNICATIONS COMMISSION, 1944, *Engineering Dept., Docket No.* 6651.
FERRELL, O. P., 1946, *Phys. Rev.*, **69,** 32.
GILLILAND, T. R., 1935, *Proc. Inst. Radio Engrs.*, **23,** 1076.
HAFSTAD, L. R. and TUVE, M. A., 1929, *Terr. Magn. Atmos. Elect.*, **34,** 39.
HARANG, L., 1942, *Geofysike Publikas.* (Oslo), **13,** No. 4 ; 1946, *Terr. Magn. Atmos. Elect.*, **51,** 581.
HARANG, L., and STOFFREGEN, W., 1938, *Nature, Lond.*, **142,** 832.
HEISING, R. A., 1928, *Proc. Inst. Radio Engrs.*, **16,** 75.
HERLOFSON, N., 1948, this Report, p. 444.
HEY, J. S., PARSONS, S. J., and STEWART, G. S., 1947, *Mon. Not. R. Astr. Soc.*, **107,** 176.
HEY, J. S., and STEWART, G. S., 1946, *Nature, Lond.*, **158,** 481 ; 1947, *Proc. Phys. Soc.*, **59,** 858.
JONES, L. F., 1933, *Proc. Inst. Radio Engrs.*, **21,** 349.
KHASTGIR, S. R., 1943, *Indian J. Phys.*, **17,** 239.
KIRBY, S. S., BERKNER, L. V., and STUART, D. M., 1934, *Bur. Stand. J. Res., Wash.*, **12,** 15.
KIRBY, S. S., GILLILAND, T. R., SMITH, N., and REYMER, S. E, 1936, *Phys. Rev.*, **50,** 258.
KIRBY, S. S., and JUDSON, E. B., 1935 a, *Bur. Stand. J. Res., Wash.*, **14,** 469; 1935 b, *Proc. Inst. Radio Engrs.*, **23,** 733.
LOVELL, A. C. B., 1947 a, *Brit. Astr. Assocn. Circ. No.* 282; 1947 b, *Nature, Lond.*, **160,** 670.
LOVELL, A. C. B., BANWELL, C. J., and CLEGG, J. A., 1947, *Mon. Not. R. Astr. Soc.*, **170,** 164.
LOVELL, A. C. B., and CLEGG, J. A., 1948, *Proc. Phys. Soc.*, **60,** 491.
LOVELL, A. C. B., CLEGG, J. A., and ELLYETT, C. D., 1947, *Nature, Lond.*, **160,** 372.
MAEDA, K., KONOMI, M., and ISAGAWA, Y., 1935, *Rep. Radio Res. Japan*, **5,** 151.
MANNING, L. A., HELLIWELL, R. A., VILLARD, O. G., and EVANS, W. E., 1946, *Phys. Rev.*, **70,** 767.
MARIS, H. B., 1929, *Terr. Magn. Atmos. Elect.*, **34,** 309.
MARTYN, D. F., and PULLEY, O. O., 1936, *Proc. Roy. Soc.*, A, **154,** 455.
MIMNO, H. R., 1934, *Nature, Lond.*, **134,** 63 ; 1937, *Rev. Mod. Phys.*, **9,** 1.
MIMNO, H. R., and WANG, P. H., 1933, *Phys. Rev.*, **43,** 769.
MINOHARA, T., and ITO, Y., 1933, *Rep. Radio Res. Japan*, **3,** 115.
MITRA, S. K., SYAM, P., and GHOSE, B. N., 1934, *Nature, Lond.*, **133,** 533.
MÖGEL, H., 1930, *Telefunken Z.*, **11,** 14 ; 1932, *Ibid.*, **13,** 32.
NAGAOKA, H., 1929, *Imp. Acad. Tokyo Proc.*, **5,** 233 ; 1931, *Inst. Phys. Chem. Res. Sci. Papers, Tokyo*, **15,** 169.
PICKARD, G. W., 1931, *Proc. Inst. Radio Engrs.*, **19,** 1166.
PIERCE, J. A., 1938, *Proc. Inst. Radio Engrs.*, **26,** 892 ; 1941, *Phys. Rev.*, **59,** 625 ; 1947, *Ibid.*, **71,** 88.
PLENDL, H., 1932, *Proc. Inst. Radio Engrs.*, **20,** 520.
PORTER, J. G., 1944, *Mon. Not. R. Astr. Soc.*, **104,** 257.
PRENTICE, J. P. M., 1948, this Report, p. 389.
PRENTICE, J. P. M., LOVELL, A. C. B., and BANWELL, C. J., 1947, *Mon. Not. R. Astr. Soc.*, **107,** 155.
QUÄCK, E., 1931, *Elekt. Nachr.-Tech.*, **8,** 46.
RANZI, I., 1932, *Nature, Lond.*, **130,** 368.
RATCLIFFE, J. A., and PAWSEY, J. L., 1933, *Proc. Camb. Phil. Soc.*, **29,** 301.

RATCLIFFE, J. A., and WHITE, E. L. C., 1933 a, *Proc. Phys. Soc.*, **45**, 399; 1933 b, *Phil. Mag.*, **16**, 125; 1934, *Proc. Phys. Soc.*, **46**, 107.
SCHAFER, J. P., and GOODALL, W. M., 1931, *Proc. Inst. Radio Engrs.*, **19**, 1434; 1932 a, *Ibid.*, **20**, 1131; 1932 b, *Ibid.*, **20**, 1941.
SKELLETT, A. M., 1931, *Phys. Rev.*, **37**, 1668; 1932, *Proc. Inst. Radio Engrs.*, **20**, 1933; 1935, *Ibid.*, **23**, 132; 1938, *Nature, Lond.*, **141**, 472.
TAYLOR, A. H., and YOUNG, L. C., 1928, *Proc. Inst. Radio Engrs.*, **16**, 561; 1929, *Ibid.*, **17**, 1491.
WATSON, FLETCHER, 1941, *Between the Planets* (Philadelphia: Blakiston Co.).
WATSON-WATT, R. A., WILKINS, A. F., and BOWEN, E. G., 1937, *Proc. Roy. Soc.*, A, **161**, 181.
WELLS, H. W., WATTS, J. M., and GEORGE, D. E., 1946, *Phys. Rev.*, **69**, 540.
WHATMAN, A. B., and HAMILTON, R. A., 1938, *Proc. Phys. Soc.*, **50**, 217.
WHIPPLE, F. L., 1943, *Rev. Mod. Phys.*, **15**, 246.
WILSON, C. T. R., 1925, *Proc. Phys. Soc.*, **37**, 32 D.

THE THEORY OF METEOR IONIZATION

By N. HERLOFSON

§1. REVIEW OF THEORIES OF METEORS

SUMMARIES of the current theories of meteors have been given by Hoffmeister (1937) and Whipple (1943). Since the views held by different authors are rather conflicting, a brief discussion of the more important theories will be given, to provide the safest possible basis for the later calculations.

The first contribution of lasting value was given by Lindemann and Dobson (1923), who ascribed the visible radiation from the meteor to evaporated atoms colliding with the air molecules, and showed that the black-body radiation from the meteor itself is negligible. The further suggestion that the meteor is largely protected against the impacts of the air molecules by an air cap with a temperature of 2000–3000° K. has been criticized on thermodynamical grounds by Sparrow (1926) and Epstein (1931), but has been maintained by Lindemann (1927). The question is of considerable interest for the study of the upper atmosphere, since this point is crucial for Lindemann and Dobson's conclusion that the atmospheric density at 100–150 km. height is not less than 10^{-9} g/cm^3. For the main part of this region such a density would exceed by a factor 100 or more the values obtained later by Appleton (1937) and others from radio investigations, and by Harang (1946) from studies of the light distribution in aurorae (cf. also a review by Bates and Massey 1946).

It will be shown that the cap does not possess the properties ascribed to it by Lindemann and Dobson, but the discrepancy does not arise directly from thermodynamical difficulties.

Assume for the moment, following Lindemann and Dobson, that the air mass intercepted may be stored as an adiabatically compressed cap in front of the meteor. This process consists essentially in a compression of the intercepted air mass, followed by its acceleration to the velocity of the meteor. The work of compression is very small compared with the kinetic energy imparted to the air cap. Since the cap and meteor remain in contact, the interaction between them may be considered as an inelastic collision, and an amount of energy, equal

to the kinetic energy imparted to the cap, is dissipated in heat in order to conserve energy and momentum.

For this reason, the assumption that the layer of gas in front of the meteor will be heated by adiabatic compression only is not permissible, and leads to a serious underestimate of the amount of heat created in the meteor and the air cap. The value deduced for the efficiency factor k, giving the fraction of the total work done which goes to heat the meteor, is much too small, of the order 0·001 to 0·01 instead of 0·1 to 1, with the result that three of the methods of Lindemann and Dobson for deriving the density of the upper atmosphere, which all depend on their estimate of k, lead to concordant values which are too large by a factor of the order of 100.

The fourth method of Lindemann and Dobson involves the relation between the dimensions of the meteor and the free path of the air molecules at the moment when an air cap is formed. A closer analysis of the collisions between the resting air molecules in front of the meteor, and air molecules which have hit the meteor surface once, indicates that an air cap may be formed when the mean free path is comparable with the dimensions of the meteor, and not with the dimensions multiplied by the ratio between the velocity of the meteor and the thermal velocities of the air molecules. Therefore, the fourth evaluation of the atmospheric density seems to involve an *underestimate* of the air density by a factor up to 100, and the agreement with the three previous methods must be a numerical coincidence.

If the last conclusion is correct, an adiabatic air cap cannot be expected to form in front of meteors even as large as some centimetres, unless they penetrate to heights well below 100 km., and for the ordinary shooting stars, whose dimensions are probably 0·01–1 cm., an air cap can hardly be supposed to form above heights of the order 50–70 km.

However, when a meteor is able to reach levels where the mean free path of the air molecules is substantially smaller than its dimensions, the physical processes undergo radical changes. The hydrodynamical equations become valid, and it is likely that a shock wave will be formed, as discussed by Epstein (1931). In a shock wave the meteor largely cuts through the air layers without impressing its own velocity on the air, and the compressions in the shock wave will be the only source of heat. In this way it may be understood qualitatively why the estimates of the atmospheric density derived from the theory of Lindemann and Dobson approach the correct values at heights lower than about 50 km.

In this connection it may be mentioned that the formation of a protecting shock wave, through which the air will pass mainly subject to temporary compressions, seems to be essential to enable small meteors of only a few centimetres diameter to penetrate the whole atmosphere without evaporating completely.

As an alternative to the Lindemann and Dobson theory of the adiabatically compressed air cap, Sparrow (1926) discussed the behaviour of meteors, subject to direct impacts of single air molecules, in the layers above the level where an air cap can be formed. Although not developed by Sparrow into a complete theory of meteors, this conception has formed the basis for all later discussions of the major number of meteors which appear as faint shooting stars in the high atmosphere.

An essential problem in the applications of this model is the study of a collision between an air molecule at rest and the surface of a solid body, for instance, composed of iron, moving with a velocity of 20 to 70 km/sec. An air molecule of mass m colliding inelastically with a meteor of velocity v will receive a kinetic energy $\frac{1}{2}mv^2$, and to conserve energy and momentum, an energy $\frac{1}{2}mv^2$ must be dissipated simultaneously. The kinetic energy of the meteor will decrease by mv^2. In the case of an *elastic* central collision, the air molecule would receive, and the meteor lose, an energy amounting to $2\,mv^2$. For the physical processes occurring at the point of impact, the situation is the same as if the meteor were at rest, and the air molecules impinged with a velocity v. In the case of oxygen or nitrogen atoms in the upper atmosphere, the kinetic energies relative to the meteor would amount to roughly from 30 upwards to 400 electron volts for the range of meteor velocities ; for diatomic molecules the energies would be doubled.

Using a very simple model to illustrate the impact of an air molecule on the atoms bound in the lattice of the solid, Sparrow estimated that a substantial fraction of the energy of the air molecules would go into heating and evaporation of the meteor body.

Although no experimental results seem to be available for air molecules of a few hundred volts impinging on solid surfaces, we may obtain some guidance from similar experiments carried out with positive ions for the study of cathode sputtering (review by Compton and Langmuir 1930), and from experiments by Voorhis and Compton (1931) on the accommodation coefficient for positive ions of noble gases.

The atoms in solids are bound with energies of a few electron volts. Since the experiments on sputtering (Compton and Langmuir 1930, figure 11) indicate that in the energy range under consideration only a fraction of an atom is evaporated for each impinging molecule, sputtering from the solid cannot be an essential factor in the energy balance. This conclusion presents serious difficulties to a suggestion by Maris (1929) for further development of Sparrow's theory.

On the other hand, the experiments on accommodation coefficients by Voorhis and Compton (1931) indicate that a major proportion (from 1/2 to 3/4) of the impinging molecules will be trapped in the metal surface, give up their energy as heat, and eventually be re-evaporated with negligible velocity relative to the solid.

Öpik (1933, 1937, 1940) has given a very detailed discussion of meteors based on these experiments. Öpik's theory is by far the most extensive contribution to the understanding of meteor impacts in the atmosphere, and its application is limited only by the present scarcity of theoretical and experimental data on the elementary processes which take place.

For a full account of Öpik's work, the reader may be referred to the original papers, which contain discussions of a number of interesting details. Much useful information and additional references to literature are also found in Whipple's review (1943).

But to provide a suitable background for the physical interpretation of radio echoes from meteors, the application of the main principles to a simplified model will be outlined in the following section.

§2. THE EVAPORATION OF A METEOR

The theory of meteor luminosity and ionization is still incomplete in some parts, although a number of observed facts are well understood. To indicate the reliable as well as the uncertain points in the theory, the impact of a spherical body of iron on the upper atmosphere will be discussed in some detail.

It may be assumed that the radius of the meteor is much smaller than the mean free path of the molecules in the atmospheric layers where the meteor evaporates ; the following calculations will show that this assumption is justified in all ordinary cases, except for brilliant fireballs. Under these conditions, the front surface of the meteor is subject to a bombardment by single air molecules, a great proportion of which are trapped in the metal surface, giving up their kinetic energy relative to the meteor as heat. The heat supplied by this process is sufficient to bring the meteor to the temperature of evaporation, and iron atoms will distil off the meteor with thermal velocities relative to the meteor surface. Relative to the surrounding atmosphere, the evaporated atoms will have the same velocity as the meteor, and light and ionization are produced by the subsequent collisions when the evaporated meteor atoms are slowed down by the air molecules.

The binding energies of the meteor atoms in the solid body amount to only a few electron volts for all common elements, so that the energy supplied by each air molecule on collision with the meteor surface is sufficient to evaporate a great number of atoms. The total air mass intercepted by the meteor before complete evaporation is therefore small compared with the mass of the meteor body, and the meteor is not appreciably retarded during its passage through the atmosphere. The action of the atmosphere on the meteor is to break it up into single atoms, which are dispersed in the air surrounding the track, and not to stop the flight of the solid body.

For a quantitative discussion of the salient features of the meteor theory, it is sufficient to consider a homogeneous iron meteor. As is evident from the careful compilation of physical data given by Öpik (1937), a variation in the composition will not introduce major differences in the general appearance of the meteor phenomena. It may also be justifiable to assume that all collisions between air molecules and the meteor surface are inelastic. The corrections to be applied when some collisions are elastic, have been discussed by Öpik (1937) and reviewed by Whipple (1943) ; the corrections are far smaller than present inaccuracies in other parts of the meteor theory, and may be excluded for the purpose of a general survey.

The evaporation process is determined as a balance between the heat supplied by the impacts of air molecules on the meteor surface, and the latent heat carried away by the evaporated atoms. As the heat supplied by inelastic collisions is equal to the kinetic energy imparted to the air molecules, we obtain

$$4\pi r^2 \rho_m l\, dr = \tfrac{1}{2}\pi r^2 \rho_a v^2 \sec\chi\, dz. \qquad \ldots\ldots(1)$$

In eq. (1), r is the radius of the meteor, ρ_m and l are density and latent heat of evaporation of the meteor material, v is the velocity of the meteor, ρ_a the density of the air, z height in the atmosphere, reckoned from some fixed level, and χ is the angle between the meteor track and the vertical.

The conservation of momentum is expressed by

$$\tfrac{4}{3}\pi r^3 \rho_m \, dv = \pi r^2 \rho_a v \sec\chi \, dz. \qquad \ldots\ldots(2)$$

The zenith angle χ is constant for the short part of the orbit traversed in the atmosphere. If the atmosphere is in hydrostatic equilibrium, equations (1) and (2) give after integration

$$r/r_\infty = \exp\left[-(v_\infty^2 - v^2)/(12\,l)\right] \qquad \ldots\ldots(3)$$

$$\frac{p\sec\chi}{r_\infty} = \frac{2g\rho_m}{3}\exp\left(-\frac{v_\infty^2}{12l}\right)\left[\mathrm{Ei}\left(\frac{v_\infty^2}{12l}\right) - \mathrm{Ei}\left(\frac{v^2}{12l}\right)\right]. \qquad \ldots\ldots(4)$$

r_∞ and v_∞ are radius and velocity of the meteor outside the atmosphere, $\mathrm{Ei}(x)$ is the integral logarithm, p is atmospheric pressure.

Eqn. (4) implies that the evaporation starts as soon as the meteor enters the atmosphere. The alternative assumption that the whole meteor is first heated to the temperature of evaporation, gives a slightly different form to eqn. (4), but the difference is not appreciable, because the latent heat of evaporation far exceeds the heat required to bring the material in the meteor to the boiling point.

The number of atoms evaporated in unit time is

$$n = -4\pi r^2 \rho_m (dr/dt)(N/A), \qquad \ldots\ldots(5)$$

where A is the atomic weight and N Avogadro's number.

Substituting eqn. (1) in eqn. (5), we obtain

$$\frac{n\sec\chi}{r_\infty^3} = \frac{\pi N}{2gHAl}\left(\frac{p\sec\chi}{r_\infty}\right)\left(\frac{r}{r_\infty}\right)^2 v^3. \qquad \ldots\ldots(6)$$

H is the atmospheric scale height. For the atmospheric pressure and the scale height we have adopted the values derived by Harang (1946) from aurorae and radio observations (cf. also Bates and Massey 1946).

Equations (3), (4) and (6) show that v, r/r_∞ and $n\sec\chi/r_\infty^3$ depend only on $p\sec\chi/r_\infty$ and on the initial velocity v_∞, if the composition of the meteor is given. In figure 1 the three functions are shown for three values of v_∞: 20 km/sec., 40 km/sec., and 60 km/sec. The intermediate value corresponds closely to the parabolic velocity at the Earth's distance from the Sun, the lower value 20 km/sec. is typical for a slow meteor overtaking the Earth from the rear, and the upper value 60 km/sec. is typical for a fast meteor coming in against the front side in the Earth's orbital motion.

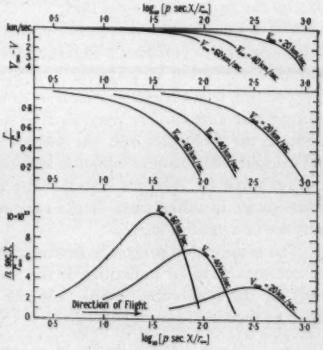


Figure 1.

It is worth noting that the velocity should not be expected to decrease materially before the meteor is practically evaporated, and that the loss in speed during the main phase is smaller for the faster meteors than for the slower. The observational evidence on this point is somewhat conflicting (Prentice 1948,

Whipple 1943), and further observations are desirable. Accurate determinations of velocity might be possible by radar methods, either by a study of the rate of growth of the echo intensity as a meteor is passing the point where the normal from the radar station intersects the meteor track, or by observing the difference in time for the first appearance of the echo at different receivers spaced some distance apart.

If the conclusion is accepted that the velocity must be practically constant, eqn. (6) shows that the rate of evaporation along the meteor track is determined as a balance between the effects of increasing pressure and decreasing meteor radius. If the velocity is nearly constant, the meteor will act as a source of homogeneous atoms shot into the atmosphere, and in the following collisions the evaporated atoms will distribute their kinetic energy in the same proportion on heat, radiation and ionization along the entire path in the air. In particular, the curve showing the rate of evaporation n, will at the same time show the distribution of light and of ionization along the meteor track.

The shape of the theoretical light-curves correspond well to the general appearance of meteors (cf. Hoffleit 1933, Foster 1942), and as shown by Whipple (1943) the values of the atmospheric scale height inferred from the length of the meteor tracks are in good agreement with accepted values from other sources of information (Harang 1946). Judging from this agreement, there is reason to accept the theory of the rate of evaporation of meteors as sufficiently reliable to give the right order of magnitude. Some details may still be in need of further study, as for instance the question to what extent the evaporating atoms shield the front surface from impacts by air molecules. But perfections in these directions may not be expected to alter the order of magnitude indicated by the simple theory.

For applications to the analysis of radio echoes from meteors, approximations to equations (3), (4) and (6) may be useful. In eqn. (4) Ei(x) may be replaced by e^x/x, and taking into account that the velocity variations are very small, it can be shown that at the point of maximum rate of evaporation the following approximations will be valid with good accuracy :

$$n_{\max} \simeq \tfrac{4}{9}(N/HA)(\tfrac{4}{3}\pi r_\infty^3 \rho_m v_\infty \cos\chi). \qquad \ldots\ldots(7)$$

$$p_{\max} \simeq \tfrac{8}{81} lg\rho_m (r_\infty / v_\infty^2) \cos\chi. \qquad \ldots\ldots(8)$$

$$r_{\max} \simeq \tfrac{2}{3} r_\infty. \qquad \ldots\ldots(9)$$

$$v_\infty - v_{\max} \simeq \frac{6l \log na + 1{\cdot}5}{v_\infty}. \qquad \ldots\ldots(10)$$

For the rate of evaporation at any point of the track, we obtain under the same conditions

$$n \simeq \tfrac{9}{4} n_{\max} (p/p_{\max})(1 - \tfrac{1}{6}(p/p_{\max}))^2. \qquad \ldots\ldots(11)$$

Eqn. (7) shows that the maximum rate of evaporation is proportional to the vertical momentum of the meteor on entering the atmosphere. The radius of a meteor may be inferred from eqn. (8) from observations of the height of the point of maximum luminosity and reasonable assumptions on its composition. Eqn. (11) shows that the shape of the evaporation curve does not depend on the initial size or the velocity of the meteor.

§3. PRODUCTION OF LIGHT AND IONS

The atoms which evaporate from the meteor are dispersed in the surrounding atmosphere. Their kinetic energies are essentially determined by the velocity of the meteor, and for iron atoms will vary from about 100 to 1000 electron volts for velocities from 20 to 60 km/sec.

The collisions with air molecules convert the kinetic energy into heat, light and ionization. The proportions in which the kinetic energy is spent on each of these three effects are controlled by the relative probabilities for transitions to the various energy levels in the colliding particles. The present knowledge about transition probabilities in collisions between heavy particles in the energy range from 100 to 1000 electron volts is very limited, and the available theoretical and experimental data can at most se ve to give an estimate of the order of magnitude of the various effects.

It is well established that the probabilities for excitation and ionization in collisions between two heavy particles are very much smaller than the collisions between an atom and an electron of comparable energy. As an example, figure 2 shows observations by Junkelmann (1937) on the excitation of the argon line $\lambda 4259$ produced by protons and deuterons (see also a review by Maurer 1939). In the case of electrons, the probability for excitation has a maximum immediately above the minimum energy required, whereas comparable probabilities are not reached before the kinetic energies exceed 10 000 electron volts when the excitation is produced by impacts of hydrogen atoms. Similar conditions hold for ionization (Sutton and Mouzon 1931) where electrons with kinetic energies of some few electron volts produce the same number of ions as heavy particles with energies of the order 10 000 electron volts. The difference in behaviour of electrons and heavy particles is due to the difference in mass, which requires a larger transfer of momentum for heavy particles than for electrons to transform a given amount of energy into excitation (cf. Massey and Smith 1933).

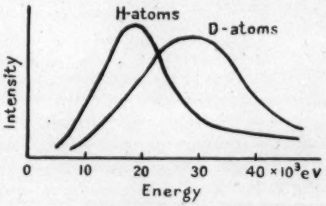


Figure 2. Comparison of the excitation of the Ar-line $\lambda 4259$ by H-atoms and D-atoms. (Junkelman, *Zeitschrift für Physik*, 1937, **107**, 561.)

On general grounds it is therefore to be expected that nearly all the kinetic energy of a meteor will be converted directly into heat, a small fraction will appear as light, and only a minute part will produce ions. This conclusion is important for any attempt to interpret radio echoes from meteors. An amount of observational material (Skellett 1932, 1935, Pierce 1941, 1947) has been discussed on the assumption that the major part of the kinetic energy of a meteor appears as ionization, and a reinterpretation would be necessary.

To calculate the various probabilities, a theory of collisions has been proposed by Öpik (1933). The theory is built on a combination of wave-mechanical considerations and a discussion of the diameters of electron orbits according to the old quantum theory. It appears likely that the cross-sections for excitation to lower potentials indicate the right order of magnitude (loc. cit. p. 32), but the probabilities for ionization must be strongly reduced by an empirical factor

(loc. cit. p. 18) to be brought into agreement with experimental data obtained by Sutton and Mouzon (1931).

The experiments by Sutton and Mouzon are performed with alkali ions colliding with atoms of noble gases, and do not give direct information on collisions between iron atoms and air molecules. On the other hand, the observed probabilities of ionization do not change in order of magnitude, except in the special cases when the masses of the two colliding particles are nearly equal, and the ionization increases strongly.

Until better theoretical and experimental data are available, the most satisfactory procedure may be to adopt for the emission of light the order of magnitude indicated by Öpik's theory, and to rely on the experimental data by Sutton and Mouzon for estimates of the probability of ionization.

The data given by Öpik (loc. cit. table XI) indicate that about 1% of the kinetic energy of a meteor with a velocity of 40 km/sec. will appear as light. The experiments by Sutton and Mouzon indicate that the probability for production of a free electron by a meteor atom before it is stopped is of the order 10^{-2}. If these rough estimates are correct, the kinetic energy of a typical meteor will be divided on heat, light and ionization in the ratios $10^4 : 10^2 : 1$.

If this estimate is combined with the calculated rate of evaporation, theoretical values can be obtained for the light and ionization produced by a meteor of given size.

Eqn. (6) or figure 1 show that for a meteor velocity of 40 km/sec., the maximum number of iron atoms evaporated per unit time is

$$n_{\max} = 7 \times 10^{23} r_\infty^3. \qquad \ldots\ldots(12)$$

The corresponding number of free electrons produced per centimetre of the meteor path is

$$a_{\max} = 10^{-2} n_{\max}/v = 2 \times 10^{15} r_\infty^3. \qquad \ldots\ldots(13)$$

The maximum energy emitted as light by the same meteor per unit time is

$$I = \tfrac{1}{2} n_{\max} (A/N) v^2 \times 10^{-2} = 5 \times 10^{12} r_\infty^3 \text{ erg/sec.} \qquad \ldots\ldots(14)$$

According to Öpik (1937), the stellar magnitude of an object at 100 km. distance emitting light of intensity I in all directions is

$$m = 24{\cdot}6 - 2{\cdot}5 \log I. \qquad \ldots\ldots(15)$$

Equations (14) and (15) give

$$m = -7{\cdot}2 - 7{\cdot}5 \log r_\infty. \qquad \ldots\ldots(16)$$

Combining the information from equations (8), (13) and (16), simultaneous values for the height of the point of maximum luminosity, the maximum number of electrons produced per unit length of the track, and the maximum stellar magnitude can be estimated for a meteor with given radius and a velocity of the order 40 km/sec. In the following table are given the calculated data for meteors of stellar magnitude +1 and +6, which are representative limits for the majority of visual meteors.

Stellar magnitude	+1	+6
Radius	0·08 cm.	0·02 cm.
No. of electrons per cm. path	10^{12}	10^{10}
Height corresponding to maximum of light and ionization	85 km.	95 km.

It should be borne in mind that the figures cannot claim to indicate more than the right order of magnitude.

Observations by Prentice, Lovell and Banwell (1947), by Lovell, Banwell and Clegg (1947) and by Lovell (1947) confirm these estimates, and suggest that the calculations are not seriously in error.

For further improvements of the theoretical calculations, better theoretical and experimental data on the collision processes between meteor atoms and air molecules are very desirable.

§4. THE STRUCTURE OF THE IONIZED COLUMN BEHIND A METEOR

The ionized column behind a meteor can only be detected by means of radio observations, and there is an intimate relationship between the study of the structure of the column and the interpretation of the radio echoes.

Electrostatic forces prevent a separation of the free electrons and the positive ions. The distribution of electrons over a cross-section of the column will therefore be determined by the combined diffusion of ions and electrons in the Earth's magnetic field, and the effective diffusion coefficient for the ionized matter will be a weighted mean of the diffusion coefficients for electrons and positive ions. In the layers where meteors most commonly occur, the air density is so small that the diffusion of electrons is markedly slower normal to the magnetic field of the Earth than in the direction of the field. Diffusion will therefore spread the ionized matter into an elliptical distribution round the track of the meteor with the minor axes of the ellipses in a direction normal to the Earth's magnetic field and the track of the meteor. If the track follows a magnetic field line, the ion distribution will have axial symmetry.

The scattering of a spherical radio wave by the ionized column behind a meteor has been discussed on very different assumptions by Pierce (1941) and by Lovell (1948). It is assumed by Pierce that the diameter of the column is considerably larger than the radio wavelength, and that the electron density is sufficiently high to give total reflection of the radiation. For frequencies in the range from 10 to 100 Mc/sec., such densities would require a much higher number of electrons per unit length of the meteor path than is indicated by the calculations discussed in the preceding section.

The calculations by Lovell follow a theory developed by Blackett and Lovell (1941). They are based on the assumptions that the incident wave is not appreciably modified by the electrons in the ionized column, and that the column is so thin that all electrons in the same cross section will scatter with the same phase. To test the validity of the assumptions, a numerical solution of the electromagnetic equations is required. Preliminary results by this method indicate that for radio echoes observed on a wavelength of 4 metres, the formula used by Lovell is valid immediately after the formation of the ionized column, provided the electron density does not exceed 10^{13} electrons per centimetre of the column. This upper limit is well above the values of 10^{10} to 10^{12} electrons per centimetre, which may be expected on theoretical grounds, and which have been observed by Prentice, Lovell and Banwell (1947) and Lovell, Banwell and Clegg (1947).

The following diffusion will introduce appreciable phase differences over a cross section of the ionized column. Rigorous calculations of the scattering from an elliptical column of ionized matter have not yet been given, but an approximate formula is valid when the ionized matter has been dispersed so much that the incident radio wave is not appreciably modified by the charges. In this case phase differences over a cross-section reduce the echo amplitude, as given by Lovell's formula, by a factor

$$\exp\left[-(16\pi^2/\lambda^2)(D_{\parallel}\cos^2\theta + D_{\perp}\sin^2\theta)t\right] = \exp\left(-t/T\right). \qquad \ldots\ldots(17)$$

λ is the radio wavelength in vacuum, $D_{\parallel}$ and $D_{\perp}$ are the diffusion coefficients for the ionized matter along and normal to the magnetic field of the Earth, θ is the angle between the direction of the Earth's magnetic field and a straight line from the radio station to the reflection point on the meteor trail, and t is the time.

If diffusion alone controlled the decay of the echo, the constant T in eqn. (17) would determine the shape of the decay curve. It is worth noting that for $\lambda = 4$ metres and with reasonable assumptions for the diffusion coefficients $D_{\parallel}$ and $D_{\perp}$, T would assume values from 0·1 to 100 sec. These values are in good accordance with durations observed by Prentice, Lovell and Banwell (1947) and Lovell, Banwell and Clegg (1947) (cf. also Lovell 1948). Eqn. (17) also suggests that for equal amplitudes, the duration should be proportional to λ^2, and that longer durations should be predominant for echoes observed in directions normal to the magnetic field of the Earth, because $D_{\parallel} > D_{\perp}$. As discussed by Lovell (1948), effects of this kind have been observed.

Experimental methods which would permit an exact determination of the height of the reflection point on the meteor trail together with a record of the time variations of radio echoes, may be expected to give information on the diffusion and recombination of ions at meteor levels.

Beside a regular decay, a number of echoes of long duration show periodic fluctuations in intensity (Lovell, Banwell and Clegg 1947, Lovell 1948). As suggested by Herlofson (cf. Lovell 1948), such fluctuations may be due to variations of the wind with height in the atmosphere which would deform the straight column left by the meteor. Observations in different latitudes can provide a decisive test for this suggestion as well as for the influence of the magnetic field of the Earth on the cross-section of the ionized column.

REFERENCES

Appleton, E. V., 1937, *Proc. Roy. Soc.* A, **162,** 472.
Bates, D. R., and Massey, H. S. W., 1946, *Proc. Roy. Soc.* A, **187,** 261.
Blackett, P. M. S., and Lovell, A. C. B., 1941, *Proc. Roy. Soc.* A, **177,** 183.
Compton, K. T., and Langmuir, I., 1930, *Rev. Mod. Phys.*, **2,** 186.
Epstein, P., 1931, *Proc. Nat. Acad. Sci.*, **17,** 532.
Foster, J. F., 1942, *Harvard Bull.*, No. 916.
Harang, L., 1946, *Terr. Magn. Atmos. Elect.*, **51,** 381.
Hoffleit, D., 1933, *Proc. Nat. Acad. Sci.*, **19,** 212.
Hoffmeister, C. F., 1937, *Die Meteore* (Berlin : Springer).
Junkelmann, R., 1937, *Z. Phys.*, **107,** 561.
Lindemann, F. A., 1927, *Astrophys. J.*, **65,** 117.
Lindemann, F. A., and Dobson, G. M. B., 1923, *Proc. Roy. Soc.* A, **102,** 411.
Lovell, A. C. B., 1947, *Nature, Lond.*, **160,** 670 ; 1948, this Report, p. 415.
Lovell, A. C. B., Banwell, C. J., and Clegg, J. A., 1947, *Mon. Not. R. Astr. Soc.*, **107,** 164.

MARIS, H. B., 1929, *Terr. Magn. Atmos. Elect.*, **34**, 309.
MASSEY, H. S. W., and SMITH, R. A., 1933, *Proc. Roy. Soc.* A, **142**, 142.
MAURER, W., 1939, *Phys. Z.*, **40**, 161.
ÖPIK, E., 1933, *Acta Univ. Tartu*, A, 26 : *Harvard Reprints* No. 100 ; 1937, *Publ. Observ. Astron. Univ. Tartu*, **24**, No. 5 ; 1940, *Mon. Not. R. Astr. Soc.*, **100**, 315.
PIERCE, J. A., 1941, *Phys. Rev.*, **59**, 625 ; 1947, *Ibid.*, **71**, 88.
PRENTICE, J. P. M., 1948, this Report, p. 389.
PRENTICE, J. P. M., LOVELL, A. C. B., and BANWELL, C. J., 1947, *Mon. Not. R. Astr. Soc.*, **107**, 155.
SKELLETT, A. M., 1932, *Proc. Inst. Radio Engrs.*, **2**, 1935 ; 1935, *Ibid.*, **23**, 132.
SPARROW, C. M., 1926, *Astrophys. J.*, **63**, 90.
SUTTON, R. M., and MOUZON, J. C., 1929, *Phys. Rev.*, **37**, 379.
VAN VOORHIS, C. C., and COMPTON, K. T., 1931, *Phys. Rev.*, **37**, 1596.
WHIPPLE, F. L., 1943, *Rev. Mod. Phys.*, **15**, 246.

REPORTS ON PROGRESS IN PHYSICS

Vol. 1 (1934). Pp. 371. Price 30*s*. (*Out of print.*)

Articles by: G. Temple, G. C. McVittie, A. Ferguson, G. F. P. Trubridge, E. G. Richardson, R. W. B. Pearse, H. Dingle, E. Gwynne Jones, C. V. Jackson, A. Hunter, T. Smith, J. S. Anderson, J. Guild, H. F. Buckley, J. H. Awbery, J. C. McLennan, G. E. Bell, I. Backhurst, W. Binks, G. Shearer, W. A. Wood, J. Thewlis, E. N. da C. Andrade, L. Hartshorn.

Vol. 2 (1935). Pp. 371. Price 30*s*. (*Out of print.*)

Articles by: A. Ferguson, G. Temple, R. Peierls, G. B. B. M. Sutherland, W. G. Penney, R. Schlapp, N. Feather, E. Lancaster-Jones, E. V. Appleton, E. G. Richardson, J. H. Awbery, L. Hartshorn, G. I. Finch, J. L. Miller, J. E. L. Robinson, H. J. H. Starks, G. E. Bell, I. Backhurst, W. Binks, W. E. Perry, G. Shearer, W. A. Wood, J. Thewlis, C. V. Jackson, R. W. B. Pearse, A. Hunter, Lotte Kellner, E. Gwynne-Jones, T. Smith, J. Guild.

Vol. 3 (1936). Pp. 394. Price 30*s*. (*Out of print.*)

Articles by: R. C. Brown, H. L. Cox, D. G. Sopwith, G. Barr, A. Fage, S. Chapman, W. C. Price, N. Feather, A. Nunn May, E. G. Richardson, G. W. C. Kaye, R. W. Powell, L. F. Bates, L. Hartshorn, D. A. Oliver, C. E. Webb, P. Vigoureux, C. E. Wynn-Williams, A. H. Wilson, G. A. Boutry, L. C. Martin, B. K. Johnson, F. Simeon, W. Swaine, W. D. Wright, W. S. Stiles, J. T. Randall, G. E. Bell, I. Backhurst, W. Binks, W. A. Wood, W. E. T. Perry, W. Jevons, M. F. Crawford, A. Hunter, R. W. B. Pearse, W. H. J. Childs.

Vol. 4 (1937). Pp. 396. Price 30*s*. (*Now available.*)

Articles by: H. Spencer Jones, E. K. Rideal, R. C. Brown, E. G. Richardson, R. W. Powell, J. H. Awbery, E. Griffiths, H. T. Flint, C. H. Collie, P. B. Moon, H. R. Robinson, L. Hartshorn, T. I. Jones, W. H. Ward, R. W. Gurney, F. Simeon, W. Jevons, W. R. S. Garton, A. Hunter, R. W. B. Pearse, E. W. Foster, Lotte Kellner, J. M. Robertson, K. Lonsdale.

Vol. 5 (1938). Pp. 445. Price 30*s*. (*Out of print.*)

Articles by: R. C. Brown, H. L. Cox, C. F. Goodeve, J. W. McBain, G. F. Mills, N. F. Mott, R. W. Gurney, H. W. Melville, A. C. D. Crommelin, D. Brunt, E. G. Richardson, P. M. Tookey Kerridge, R. W. Powell, L. C. Martin, W. M. Hampton, W. N. Wheat, E. W. H. Selwyn, W. D. Wright, W. Jevons, A. G. Shenstone, M. F. Crawford, J. Stuart Foster, A. Hunter, R. W. B. Pearse, H. W. B. Skinner, W. V. Mayneord, L. Hartshorn, N. F. Astbury, F. C. Champion, W. Heitler, A. Nunn May, H. T. Flint, A. W. Barton.

Vol. 6 (1939). Pp. 434. Price 30*s*. (*Out of print.*)

Articles by: H. A. Bethe, R. E. Marshak, H. J. Walke, H. C. Urey, G. W. C. Kaye, G. E. Bell, W. Binks, W. E. Perry, W. B. Mann, W. Jevons, A. Hunter, H. E. White, F. H. Crawford, J. T. Randall, M. H. F. Wilkins, N. F. Mott, W. G. Penney, E. G. Richardson, A. Fage, H. Jones, R. W. Powell, H. Pelzer, L. C. Jackson, J. M. Robertson, W. H. Taylor, H. Lipson, E. T. Goodwin, L. Hartshorn, R. S. Rivlin, N. F. Astbury, H. Fröhlich, D. Owen.

Vol. 7 (1940). Pp. 362. Price 30*s*. (*Out of print.*)

Articles by: E. G. Richardson, W. D. Wright, H. H. Nielsen, N. Feather, R. Peierls, O. Klemperer, H. W. Lee, A. Hunter, A. D. Thackeray, R. C. Brown, J. A. Beattie, W. H. Stockmayer, V. J. Francis, H. G. Jenkins, J. K. Roberts, F. Eirich, A. Ferguson.

Vol. 8 (1941). Pp. 372. Price 30*s*. (*Out of print.*)

Articles by: M. Born, M. Blackman, J. W. Beams, A. G. Gaydon, R. Schnurmann, R. T. Birge, V. K. Zworykin, R. E. Shelby, A. Hunter, Lotte Kellner, G. R. Harrison, R. S. Mulliken, C. A. Rieke, E. P. Wigner, E. Feenberg, J. M. Meek.

Vol. 9 (1942–43). Pp. 353. Price 30*s*. (*Out of print.*)

Articles by: S. Chapman, A. Hunter, W. C. Price, G. B. B. M. Sutherland, G. S. Callender, T. G. Cowling, R. W. B. Pearse, J. Sayers, H. S. W. Massey, D. R. Bates, C. H. Bamford, L. R. G. Treloar, J. D. Craggs, A. E. Alexander, R. O. Jenkins, J. McG. Bruckshaw, E. G. Richardson, K. Lonsdale, M. Born, F. I. G. Rawlins.

Vol. 10 (1944–45). Pp. 442. Price 30*s*. (*Now available.*)

Articles by: W. F. G. Swann, H. Jeffreys, H. P. Rooksby, E. G. Richardson, P. Koets, A. B. D. Cassie, J. G. Martindale, L. Marton, H. J. Bhabha, A. F. C. Pollard, J. M. Meek, F. R. Perry, K. Mendelssohn, J. Cabannes, J. D. Bernal.

Published by

The Physical Society, 1 Lowther Gardens, Prince Consort Road, S.W.7

Orders with remittance should be sent to the above address.